(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 January 2019 (17.01.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/010564 A1

(51) International Patent Classification:
C12Q 1/6809 (2018.01) *C12Q 1/6886* (2018.01)
C12Q 1/68 (2018.01) *G06F 19/22* (2011.01)

(21) International Application Number:
PCT/CA2018/000141

(22) International Filing Date:
11 July 2018 (11.07.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/531,527 12 July 2017 (12.07.2017) US

(71) Applicant: **UNIVERSITY HEALTH NETWORK** [CA/CA]; 190 Elizabeth Street, R. Fraser Elliott Building, Room 1S-417, Toronto, Ontario M5G 2C4 (CA).

(72) Inventors: **DE CARVALHO, Daniel Diniz**; 31 Drummondville Lane, Toronto, ON M4E 3X7 (CA). **BRATMAN, Scott Victor**; 44 Soho Square, Toronto, ON M5T 2Z4 (CA). **SINGHANIA, Rajat**; 100 Wellesley St East, Apt 806, Toronto, ON M4Y1H5 (CA). **CHAKRAVARTHY, Ankur Ravinarayana**; #1414, 280 Wellesley Street East, Toronto, ON M4X 1G7 (CA).

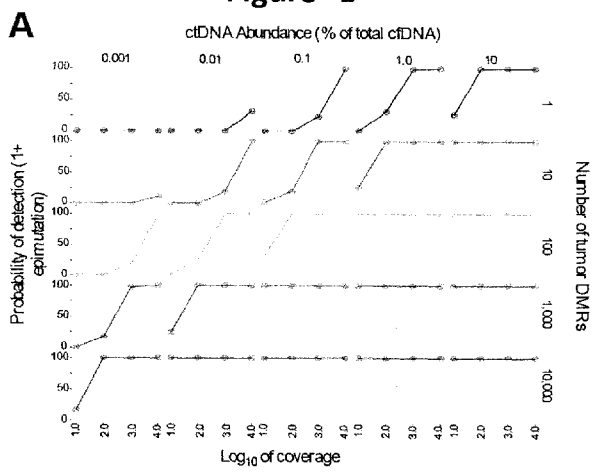
(74) Agent: **CHIU, Jung-kay**; Norton Rose Fulbright Canada LLP /S.E.N.C.R.L, s.r.l, 1, Place Ville Marie, Suite 2500, Montreal, Quebec H3B 1R1 (CA).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: CANCER DETECTION AND CLASSIFICATION USING METHYLOME ANALYSIS

Figure 1



(57) Abstract: There is described herein a method of detecting the presence of DNA from cancer cells in a subject comprising: providing a sample of cell-free DNA from a subject; subjecting the sample to library preparation to permit subsequent sequencing of the cell-free methylated DNA; adding a first amount of filler DNA to the sample, wherein at least a portion of the filler DNA is methylated, then optionally denaturing the sample; capturing cell-free methylated DNA using a binder selective for methylated polynucleotides; sequencing the captured cell-free methylated DNA; comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals and from individuals with distinct cancer types and subtypes; identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals.

WO 2019/010564 A1

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

CANCER DETECTION AND CLASSIFICATION USING METHYLOME ANALYSIS

RELATED APPLICATIONS

This application claims priority to U.S. Provisional Patent Application No. 62/531527 filed July 12, 2017, which is hereby incorporated by reference.

FIELD OF THE INVENTION

The invention relates to cancer detection and classification and more particularly to the use of methylome analysis for the same.

BACKGROUND OF THE INVENTION

The use of circulating cell-free DNA (cfDNA) as a source of biomarkers is rapidly gaining momentum in oncology[1]. Use of DNA methylation mapping of cfDNA as a biomarker could have a significant impact in the field of liquid biopsy, as it could allow for the identification of the tissue-of-origin[2], allow for cancer type and subtype classification, and stratify cancer patients in a minimally invasive fashion[3]. Furthermore, using genome-wide DNA methylation mapping of cfDNA could overcome a critical sensitivity problem in detecting circulating tumor DNA (ctDNA) in patients with early-stage cancer with no radiographic evidence of disease. Existing ctDNA detection methods are based on sequencing mutations and have limited sensitivity in part due to the limited number of recurrent mutations available to distinguish between tumor and normal circulating cfDNA[4, 5]. On the other hand, genome-wide DNA methylation mapping leverages large numbers of epigenetic alterations that may be used to distinguish circulating tumor DNA (ctDNA) from normal circulating cell-free DNA (cfDNA). For example, some tumor types, such as ependymomas, can have extensive DNA methylation aberrations without any significant recurrent somatic mutations[6].

Certain methods of capturing cell-free methylated DNA are described in WO 2017/190215, which is incorporated by reference.

SUMMARY OF THE INVENTION

In an aspect, there is provided a method of detecting the presence of DNA from cancer cells in a subject comprising: providing a sample of cell-free DNA from a subject; subjecting the sample to library preparation to permit subsequent sequencing of the cell-free methylated DNA; adding a first amount of filler DNA to the sample, wherein at least a portion of the filler DNA is methylated, then optionally denaturing the sample; capturing cell-free methylated DNA using a binder selective for methylated polynucleotides; sequencing the captured cell-free methylated DNA; comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals.

In an aspect, there is provided a method of detecting the presence of DNA from cancer cells and identifying a cancer subtype, the method comprising: receiving sequencing data of cell-free methylated DNA from a subject sample; comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals; and if DNA from cancer cells is identified, further identifying the cancer cell tissue of origin and cancer subtype based on the comparison.

In an aspect, there is provided a computer-implemented method of detecting the presence of DNA from cancer cells and identifying a cancer subtype, the method comprising: receiving, at least one processor, sequencing data of cell-free methylated DNA from a subject sample; comparing, at the at least one processor, the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identifying, at the at least one processor, the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNA sequences from cancerous individuals and if DNA from cancer cells is identified, further identifying the cancer cell tissue of origin and cancer subtype based on the comparison.

In an aspect, there is provided a computer program product for use in conjunction with a general-purpose computer having a processor and a memory connected to the processor, the computer program product comprising a computer readable storage medium having a computer mechanism encoded thereon, wherein the computer program mechanism may be loaded into the memory of the computer and cause the computer to carry out the method described herein.

In an aspect, there is provided a computer readable medium having stored thereon a data structure for storing the computer program product described herein.

In an aspect, there is provided a device for detecting the presence of DNA from cancer cells and identifying a cancer subtype, the device comprising: at least one processor; and electronic memory in communication with the at one processor, the electronic memory storing processor-executable code that, when executed at the at least one processor, causes the at least one processor to: receive sequencing data of cell-free methylated DNA from a subject sample; compare the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identify the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNA sequences from cancerous individuals and if DNA from cancer cells is identified, further identify the cancer cell tissue of origin and cancer subtype based on the comparison.

In an aspect, there is provided a method of detecting the presence of DNA from cancer cells and determining the location of the cancer from which the cancer cells arose from two or more possible organs, the method comprising: providing a sample of cell-free DNA from a subject; capturing cell-free methylated DNA from said sample, using a binder selective for methylated polynucleotides; sequencing the captured cell-free methylated DNA; comparing the sequence patterns of the captured cell-free methylated DNA to DNAs sequence patterns of two or more population(s) of control individuals, each of said two or more populations having localized cancer in a different organ; determining as to which organ the cancer cells arose on the basis of a statistically significant similarity between the pattern of methylation of the cell-free DNA and one of said two or more populations.

BRIEF DESCRIPTION OF FIGURES

These and other features of the preferred embodiments of the invention will become more apparent in the following detailed description in which reference is made to the appended drawings wherein:

Figure 1 shows methylome analysis of cfDNA is a highly sensitive approach to enrich and detect ctDNA in low amounts of input DNA. **A)** Computer Simulation of the probability to detect at least one epimutation as a function of the concentration of ctDNA (*columns*), number of DMRs being investigated (*rows*), and the sequencing depth (x-axis). **B)** Genome-wide Pearson correlation between DNA methylation signal for 1 to 100 ng of input DNA from HCT116 cell line fragmented to mimic plasma cfDNA. Each concentration has two biological replicates. **C)** DNA methylation profile obtained from cfMeDIP-seq from different concentrations of input DNA from HCT116 (Green Tracks) plus RRBS (Reduced Representation Bisulfite Sequencing) HCT116 data obtained from ENCODE (ENCSR000DFS) and WGBS (Whole-Genome Bisulfite Sequencing) HCT116 data obtained from GEO (GSM1465024). For the heatmap (RRBS track), yellow means methylated, blue means unmethylated and gray means no coverage. **D-E**) Serial dilution of the CRC cell line HCT116 into the Multiple Myeloma (MM) cell line MM1.S. cfMeDIP-seq was performed in pure HCT116 DNA (100% CRC), pure MM1.S DNA (100% MM) and 10%, 1%, 0.1%, 0.01%, and 0.001% CRC DNA diluted into MM DNA. All DNA was fragmented to mimic plasma cfDNA. We observed an almost perfect linear correlation (r^2=0.99, $p<0.0001$) between the observed versus expected (**D**) numbers of DMRs and (**E**) the DNA methylation signal (in RPKM) within those DMRs. **F**) In the same dilution series, known somatic mutations are only detectable at 1/100 allele fraction by ultra-deep (>10,000X) targeted sequencing, above the background sequencer and polymerase error rate. Shown are the fractions of reads containing each base or an insertion/deletion at the site of each mutation in the CRC cell line. **G)** Frequency of ctDNA (human) as a percentage of total cfDNA (human + mice) in the plasma of mice harboring patient-derived xenograft (PDX) from two colorectal cancer patients.

Figure 2 shows the methylome analysis of plasma cfDNA allows tumor classification. **A)** Schematic demonstrating the approach of machine learning classifier construction for cancer classification. **B)** Heatmap of DMRs contained within the multi-class elastic net machine learning classifiers. The classifiers were trained on plasma DNA samples

from healthy donors (n=24), lung cancer (n=25), breast cancer (n=25), colorectal cancer (n=23), acute myelogenous leukemia (AML) (n=28), and glioblasatoma multiforme (GBM) (n=71). Hierarchical clustering method: Ward. **C)** 2D visualizations by tSNE (t-Distributed Stochastic Neighbor Embedding) of the cancer-type associated DMRs identified in 10% or 25% of models. **D)** Performance metrics for the plasma cfDNA methylation-based multi-cancer classifier. Area under the receiver operator curve (auROC) shown on the y-axis for each cancer type and healthy donors following 50-fold generation of elastic net machine learning classifiers.

Figure 3 shows validation of the multi-cancer classifier on independent cohorts. **A)** ROC curves are shown for independent validation of the multi-cancer classifier on cohorts of lung cancer (LUC) (n=55 LUC vs n=97 other), AML (n=35 AML vs n=117 other), and healthy donors (n=62 healthy donors vs n=90 other). **B)** ROC curves are shown for independent validation of the multi-cancer classifier on early stage LUC (n=32 stage I-II LUC vs n=97 other) and late stage LUC (n=23 stage III-IV LUC vs n=97 other).

Figure 4 shows the methylome analysis of plasma cfDNA allows tumor subtype classification. **A)** 2D visualizations by tSNE (t-Distributed Stochastic Neighbor Embedding) of cancer subtype associated DMRs. Breast cancer subtypes show ability to distinguish between patients harboring tumors with distinct gene expression pattern and transcription factor activity (ER status) as well as distinct tumor copy number aberrations (HER2 status). AML subtypes show ability to distinguish between patients harboring tumors with distinct rearrangements (FLT3 status). Glioblastoma multiforme (GBM) subtypes show ability to distinguish between patients harboring tumors with distinct point mutations (IDH gene mutational status). Lung cancer subtypes show ability to distinguish between patients harboring tumors with distinct histologies that have prognostic and therapeutic implications (adenocarcinoma vs. squamous carcinoma vs. small cell carcinoma). **B)** Heatmap showing the top DMRs that allow accurate discrimination of the three breast cancer subtypes in breast cancer plasma samples. **C)** Heatmap showing the top DMRs that allow accurate discrimination of the FLT3-ITD status in AML patient plasma samples. **D)** Heatmap showing the top DMRs that allow accurate discrimination of the IDH gene mutational status in glioblastoma multiforme (GBM) patient plasma samples. **E)** Heatmap showing the top DMRs that allow accurate discrimination of the three lung cancer histologies in lung cancer plasma samples.

Figure 5 shows a suitable configured computer device, and associated communications networks, devices, software and firmware to provide a platform for enabling one or more embodiments as described herein.

Figure 6 shows sequencing saturation analysis and quality controls. **A)** The figure shows the results of the saturation analysis from the Bioconductor package MEDIPS analyzing cfMeDIP-seq data from each replicate for each input concentration from the HCT116 DNA fragmented to mimic plasma cfDNA. **B)** The protocol was tested in two replicates of four starting DNA concentrations (100, 10, 5, and 1 ng) of HCT116 cell line. Specificity of the reaction was calculated using methylated and unmethylated spiked-in A. thaliana DNA. Fold enrichment ratio was calculated using genomic regions of the fragmented HCT116 DNA (Primers for methylated testis-specific H2B, TSH2B0 and unmethylated human DNA region (GAPDH promoter)). The horizontal dotted line indicates a fold-enrichment ratio threshold of 25. Error bars represent ± 1 s.e.m. **C)** CpG Enrichment Scores of the sequenced samples show a robust enrichment of CpGs within the genomic regions from the immunoprecipitated samples compared to the input control. The CpG Enrichment Score was obtained by dividing the relative frequency of CpGs of the regions by the relative frequency of CpGs of the human genome. Error bars represent ± 1 s.e.m.

DETAILED DESCRIPTION

In the following description, numerous specific details are set forth to provide a thorough understanding of the invention. However, it is understood that the invention may be practiced without these specific details.

DNA methylation profiles are cell-type specific and are disrupted in cancer. Using a robust and sensitive method designed for methylome analysis of minute amounts of circulating cell-free DNA (cfDNA), we identified thousands of Differentially Methylated Regions (DMRs) that distinguish multiple tumor types from each other and from healthy individuals. Methylome analysis of cfDNA is highly sensitive and suitable for detecting circulating tumor DNA (ctDNA) in early stage patients. A machine-learning derived classifier using cfDNA methylomes was able to correctly classify 196 plasma samples from patients with 5 cancer types and healthy donors based on cross-validation. In an independent validation, using the same DMRs identified in the plasma

cfDNA, the classifier was able to correctly classify AML, lung cancer, and healthy donors, as well as both early and late stage lung cancer. Therefore, methylome analysis of cfDNA can be used for non-invasive early stage detection of ctDNA and robustly classify cancer types.

In an aspect, there is provided a method of detecting the presence of DNA from cancer cells in a subject comprising: providing a sample of cell-free DNA from a subject; subjecting the sample to library preparation to permit subsequent sequencing of the cell-free methylated DNA; adding a first amount of filler DNA to the sample, wherein at least a portion of the filler DNA is methylated, then optionally denaturing the sample; capturing cell-free methylated DNA using a binder selective for methylated polynucleotides; sequencing the captured cell-free methylated DNA; comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals.

Applicant's co-owned applications U.S. Provisional Patent Application No. 62/331,070 filed on May 3, 2016 and International Patent Application No. PCT/CA2017/000108 filed on May 3, 2017 describe method for capturing cell-free methylated DNA and are incorporated herein by reference.

Cancer has been traditionally classified by tissue of origin – for instance, colorectal cancer, breast cancer, lung cancer, etc. In the modern practice of clinical oncology, it is becoming increasingly important to be able to distinguish subtypes of cancer by various molecular, developmental, and functional underpinnings. Therapeutic decisions often hinge on the precise subtype of cancer, and it may be necessary for clinicians to identify the subtype prior to initiation of therapy. Examples of cancer subtyping that may influence therapeutic decisions include (but are not limited to) stage (e.g., early stage lung cancer treated with surgery vs late stage lung cancer treated with chemotherapy), histology (e.g., small cell carcinoma vs adenocarcinoma vs squamous cell carcinoma in lung cancer), gene expression pattern or transcription factor activity (e.g., ER status in breast cancer), copy number aberrations (e.g., HER2 status in breast cancer), specific rearrangements (e.g., FLT3 in AML), specific gene

point mutational status (e.g., IDH gene point mutations), and DNA methylation patterns (e.g., MGMT gene promoter methylation in brain cancer).

The methods described herein are applicable to a wide variety of cancers, including but not limited to adrenal cancer, anal cancer, bile duct cancer, bladder cancer, bone cancer, brain/cns tumors, breast cancer, castleman disease, cervical cancer, colon/rectum cancer, endometrial cancer, esophagus cancer, ewing family of tumors, eye cancer, gallbladder cancer, gastrointestinal carcinoid tumors, gastrointestinal stromal tumor (gist), gestational trophoblastic disease, hodgkin disease, kaposi sarcoma, kidney cancer, laryngeal and hypopharyngeal cancer, leukemia (acute lymphocytic, acute myeloid, chronic lymphocytic, chronic myeloid, chronic myelomonocytic), liver cancer, lung cancer (non-small cell, small cell, lung carcinoid tumor), lymphoma, lymphoma of the skin, malignant mesothelioma, multiple myeloma, myelodysplastic syndrome, nasal cavity and paranasal sinus cancer, nasopharyngeal cancer, neuroblastoma, non-hodgkin lymphoma, oral cavity and oropharyngeal cancer, osteosarcoma, ovarian cancer, penile cancer, pituitary tumors, prostate cancer, retinoblastoma, rhabdomyosarcoma, salivary gland cancer, sarcoma - adult soft tissue cancer, skin cancer (basal and squamous cell, melanoma, merkel cell), small intestine cancer, stomach cancer, testicular cancer, thymus cancer, thyroid cancer, uterine sarcoma, vaginal cancer, vulvar cancer, waldenstrom macroglobulinemia, wilms tumor.

Various sequencing techniques are known to the person skilled in the art, such as polymerase chain reaction (PCR) followed by Sanger sequencing. Also available are next-generation sequencing (NGS) techniques, also known as high-throughput sequencing, which includes various sequencing technologies including: Illumina (Solexa) sequencing, Roche 454 sequencing, Ion torrent: Proton / PGM sequencing, SOLiD sequencing. NGS allow for the sequencing of DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing. In some embodiments, said sequencing is optimized for short read sequencing.

The term "subject" as used herein refers to any member of the animal kingdom, preferably a human being and most preferably a human being that has, has had, or is suspected of having prostate cancer.

Cell-free methylated DNA is DNA that is circulating freely in the blood stream, and are methylated at various known regions of the DNA. Samples, for example, plasma

samples can be taken to analyze cell-free methylated DNA. Accordingly, in some embodiments, the sample is the subject's blood or plasma.

As used herein, "library preparation" includes list end-repair, A-tailing, adapter ligation, or any other preparation performed on the cell free DNA to permit subsequent sequencing of DNA.

As used herein, "filler DNA" can be noncoding DNA or it can consist of amplicons.

DNA samples may be denatured, for example, using sufficient heat.

In some embodiments, the comparison step is based on fit using a statistical classifier. Statistical classifiers using DNA methylation data can be used for assigning a sample to a particular disease state, such as cancer type or subtype. For the purpose of cancer type or subtype classification, a classifier would consist of one or more DNA methylation variables (i.e., features) within a statistical model, and the output of the statistical model would have one or more threshold values to distinguish between distinct disease states. The particular feature(s) and threshold value(s) that are used in the statistical classifier can be derived from prior knowledge of the cancer types or subtypes, from prior knowledge of the features that are likely to be most informative, from machine learning, or from a combination of two or more of these approaches.

In some embodiments, the classifier is machine learning-derived. Preferably, the classifier is an elastic net classifier, lasso, support vector machine, random forest, or neural network.

The genomic space that is analyzed can be genome-wide, or preferably restricted to regulatory regions (i.e., FANTOM5 enhancers, CpG Islands, CpG shores and CpG Shelves).

Preferably, the percentage of spike-in methylated DNA recovered is included as a covariate to control for pulldown efficiency variation.

For a classifier capable of distinguishing multiple cancer types (or subtypes) from one another, the classifier would preferably consist of differentially methylated regions from pairwise comparisons of each type (or subtype) of interest.

In some embodiments, the control cell-free methylated DNAs sequences from healthy and cancerous individuals are comprised in a database of Differentially Methylated Regions (DMRs) between healthy and cancerous individuals.

In some embodiments, the control cell-free methylated DNA sequences from healthy and cancerous individuals are limited to those control cell-free methylated DNA sequences which are differentially methylated as between healthy and cancerous individuals in DNA derived from cell-free DNA from bodily fluids, such as from blood serum, cerebral spinal fluid, urine stool, sputum, pleural fluid, ascites, tears, sweat, pap smear fluid, endoscopy brushings fluid, ..etc., preferably from blood plasma.

In some embodiments, the sample has less than 100 ng, 75 ng, or 50 ng of cell-free DNA.

In some embodiments, the first amount of filler DNA comprises about 5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% methylated filler DNA with remainder being unmethylated filler DNA, and preferably between 5% and 50%, between 10%-40%, or between 15%-30% methylated filler DNA.

In some embodiments, the first amount of filler DNA is from 20 ng to 100 ng, preferably 30 ng to 100 ng, more preferably 50 ng to 100 ng.

In some embodiments, the cell-free DNA from the sample and the first amount of filler DNA together comprises at least 50 ng of total DNA, preferably at least 100 ng of total DNA.

In some embodiments, he filler DNA is 50 bp to 800 bp long, preferably 100 bp to 600 bp long, and more preferably 200 bp to 600 bp long.

In some embodiments, the filler DNA is double stranded. The filler DNA is double stranded. For example, the filler DNA can be junk DNA. The filler DNA may also be endogenous or exogenous DNA. For example, the filler DNA is non-human DNA, and in preferred embodiments, λ DNA. As used herein, "λ DNA" refers to *Enterobacteria phage* λ DNA. In some embodiments, the filler DNA has no alignment to human DNA.

In some embodiments, the binder is a protein comprising a Methyl-CpG-binding domain. One such exemplary protein is MBD2 protein. As used herein, "Methyl-CpG-binding domain (MBD)" refers to certain domains of proteins and enzymes that is

approximately 70 residues long and binds to DNA that contains one or more symmetrically methylated CpGs. The MBD of MeCP2, MBD1, MBD2, MBD4 and BAZ2 mediates binding to DNA, and in cases of MeCP2, MBD1 and MBD2, preferentially to methylated CpG. Human proteins MECP2, MBD1, MBD2, MBD3, and MBD4 comprise a family of nuclear proteins related by the presence in each of a methyl-CpG-binding domain (MBD). Each of these proteins, with the exception of MBD3, is capable of binding specifically to methylated DNA.

In other embodiments, the binder is an antibody and capturing cell-free methylated DNA comprises immunoprecipitating the cell-free methylated DNA using the antibody. As used herein, "immunoprecipitation" refers a technique of precipitating an antigen (such as polypeptides and nucleotides) out of solution using an antibody that specifically binds to that particular antigen. This process can be used to isolate and concentrate a particular protein or DNA from a sample and requires that the antibody be coupled to a solid substrate at some point in the procedure. The solid substrate includes for examples beads, such as magnetic beads. Other types of beads and solid substrates are known in the art.

One exemplary antibody is 5-MeC antibody. For the immunoprecipitation procedure, in some embodiments at least 0.05 µg of the antibody is added to the sample; while in more preferred embodiments at least 0.16 µg of the antibody is added to the sample. To confirm the immunoprecipitation reaction, in some embodiments the method described herein further comprises the step of adding a second amount of control DNA to the sample.

In some embodiments, the method further comprises the step of adding a second amount of control DNA to the sample for confirming the immunoprecipitation reaction.

As used herein, the "control" may comprise both positive and negative control, or at least a positive control.

In some embodiments, the method further comprises the step of adding a second amount of control DNA to the sample for confirming the capture of cell-free methylated DNA.

In some embodiments, identifying the presence of DNA from cancer cells further includes identifying the cancer cell tissue of origin.

In some instances, tumor tissue sampling may be challenging or carry significant risks, in which case diagnosing and/or subtyping the cancer without the need for tumor tissue sampling may be desired. For example, lung tumor tissue sampling may require invasive procedures such as mediastinoscopy, thoracotomy, or percutaneous needle biopsy; these procedures may result in a need for hospitalization, chest tube, mechanical ventilation, antibiotics, or other medical interventions. Some individuals may not undergo the invasive procedures needed for tumor tissue sampling either because of medical comorbidities or due to preference. In some instances, the actual procedure for tumor tissue procurement may depend on the suspected cancer subtype. In other instances, cancer subtype may evolve over time within the same individual; serial assessment with invasive tumor tissue sampling procedures is often impractical and not well tolerated by patients. Thus, non-invasive cancer subtyping via blood test could have many advantageous applications in the practice of clinical oncology.

Accordingly, in some embodiments, identifying the cancer cell tissue of origin further includes identifying a cancer subtype. Preferably, the cancer subtype differentiates the cancer based on stage (e.g., early stage lung cancer treated with surgery vs late stage lung cancer treated with chemotherapy), histology (e.g., small cell carcinoma vs adenocarcinoma vs squamous cell carcinoma in lung cancer), gene expression pattern or transcription factor activity (e.g., ER status in breast cancer), copy number aberrations (e.g., HER2 status in breast cancer), specific rearrangements (e.g., FLT3 in AML), specific gene point mutational status (e.g., IDH gene point mutations), and DNA methylation patterns (e.g., MGMT gene promoter methylation in brain cancer).

In some embodiments, comparison in step (f) is carried out genome-wide.

In other embodiments, the comparison in step (f) is restricted from genome-wide to specific regulatory regions, such as, but not limited to, FANTOM5 enhancers, CpG Islands, CpG shores, CpG Shelves, or any combination of the foregoing.

In some embodiments, certain steps are carried out by a computer processor.

In an aspect, there is provided a method of detecting the presence of DNA from cancer cells and identifying a cancer subtype, the method comprising: receiving sequencing data of cell-free methylated DNA from a subject sample; comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences

from healthy and cancerous individuals; identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals; and if DNA from cancer cells is identified, further identifying the cancer cell tissue of origin and cancer subtype based on the comparison step.

In an aspect, there is provided a method of detecting the presence of DNA from cancer cells and determining the location of the cancer from which the cancer cells arose from two or more possible organs, the method comprising: providing a sample of cell-free DNA from a subject; capturing cell-free methylated DNA from said sample, using a binder selective for methylated polynucleotides; sequencing the captured cell-free methylated DNA; comparing the sequence patterns of the captured cell-free methylated DNA to DNAs sequence patterns of two or more population(s) of control individuals, each of said two or more populations having localized cancer in a different organ; determining as to which organ the cancer cells arose on the basis of a statistically significant similarity between the pattern of methylation of the cell-free DNA and one of said two or more populations.

The present system and method may be practiced in various embodiments. A suitably configured computer device, and associated communications networks, devices, software and firmware may provide a platform for enabling one or more embodiments as described above. By way of example, **Figure 5** shows a generic computer device 100 that may include a central processing unit ("CPU") 102 connected to a storage unit 104 and to a random access memory 106. The CPU 102 may process an operating system 101, application program 103, and data 123. The operating system 101, application program 103, and data 123 may be stored in storage unit 104 and loaded into memory 106, as may be required. Computer device 100 may further include a graphics processing unit (GPU) 122 which is operatively connected to CPU 102 and to memory 106 to offload intensive image processing calculations from CPU 102 and run these calculations in parallel with CPU 102. An operator 107 may interact with the computer device 100 using a video display 108 connected by a video interface 105, and various input/output devices such as a keyboard 115, mouse 112, and disk drive or solid state drive 114 connected by an I/O interface 109. In known manner, the mouse 112 may be configured to control movement of a cursor in the video display 108, and to operate various graphical user interface (GUI) controls appearing in the video display 108 with a mouse button. The disk drive or solid state drive 114 may be

configured to accept computer readable media 116. The computer device 100 may form part of a network via a network interface 111, allowing the computer device 100 to communicate with other suitably configured data processing systems (not shown). One or more different types of sensors 135 may be used to receive input from various sources.

The present system and method may be practiced on virtually any manner of computer device including a desktop computer, laptop computer, tablet computer or wireless handheld. The present system and method may also be implemented as a computer-readable/useable medium that includes computer program code to enable one or more computer devices to implement each of the various process steps in a method in accordance with the present invention. In case of more than computer devices performing the entire operation, the computer devices are networked to distribute the various steps of the operation. It is understood that the terms computer-readable medium or computer useable medium comprises one or more of any type of physical embodiment of the program code. In particular, the computer-readable/useable medium can comprise program code embodied on one or more portable storage articles of manufacture (e.g. an optical disc, a magnetic disk, a tape, etc.), on one or more data storage portioned of a computing device, such as memory associated with a computer and/or a storage system.

In an aspect, there is provided a computer-implemented method of detecting the presence of DNA from cancer cells and identifying a cancer subtype, the method comprising: receiving, at least one processor, sequencing data of cell-free methylated DNA from a subject sample; comparing, at the at least one processor, the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identifying, at the at least one processor, the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals and if DNA from cancer cells is identified, further identifying the cancer cell tissue of origin and cancer subtype based on the comparison step;

In an aspect, there is provided a computer program product for use in conjunction with a general-purpose computer having a processor and a memory connected to the processor, the computer program product comprising a computer readable storage

medium having a computer mechanism encoded thereon, wherein the computer program mechanism may be loaded into the memory of the computer and cause the computer to carry out the method described herein.

In an aspect, there is provided a computer readable medium having stored thereon a data structure for storing the computer program product described herein.

In an aspect, there is provided a device for detecting the presence of DNA from cancer cells and identifying a cancer subtype, the device comprising: at least one processor; and electronic memory in communication with the at one processor, the electronic memory storing processor-executable code that, when executed at the at least one processor, causes the at least one processor to: receive sequencing data of cell-free methylated DNA from a subject sample; compare the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals; identify the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals and if DNA from cancer cells from is identified, further identify the cancer cell tissue of origin and cancer subtype based on the comparison step.

As used herein, "processor" may be any type of processor, such as, for example, any type of general-purpose microprocessor or microcontroller (e.g., an Intel™ x86, PowerPC™, ARM™ processor, or the like), a digital signal processing (DSP) processor, an integrated circuit, a field programmable gate array (FPGA), or any combination thereof.

As used herein "memory" may include a suitable combination of any type of computer memory that is located either internally or externally such as, for example, random-access memory (RAM), read-only memory (ROM), compact disc read-only memory (CDROM), electro-optical memory, magneto-optical memory, erasable programmable read-only memory (EPROM), and electrically-erasable programmable read-only memory (EEPROM), or the like. Portions of memory 102 may be organized using a conventional filesystem, controlled and administered by an operating system governing overall operation of a device.

As used herein, "computer readable storage medium" (also referred to as a machine-readable medium, a processor-readable medium, or a computer usable medium

having a computer-readable program code embodied therein) is a medium capable of storing data in a format readable by a computer or machine. The machine-readable medium can be any suitable tangible, non-transitory medium, including magnetic, optical, or electrical storage medium including a diskette, compact disk read only memory (CD-ROM), memory device (volatile or non-volatile), or similar storage mechanism. The computer readable storage medium can contain various sets of instructions, code sequences, configuration information, or other data, which, when executed, cause a processor to perform steps in a method according to an embodiment of the disclosure. Those of ordinary skill in the art will appreciate that other instructions and operations necessary to implement the described implementations can also be stored on the computer readable storage medium. The instructions stored on the computer readable storage medium can be executed by a processor or other suitable processing device, and can interface with circuitry to perform the described tasks.

As used herein, "data structure" a particular way of organizing data in a computer so that it can be used efficiently. Data structures can implement one or more particular abstract data types (ADT), which specify the operations that can be performed on a data structure and the computational complexity of those operations. In comparison, a data structure is a concrete implementation of the specification provided by an ADT.

The advantages of the present invention are further illustrated by the following examples. The examples and their particular details set forth herein are presented for illustration only and should not be construed as a limitation on the claims of the present invention.

EXAMPLES

Methods and Materials

Donor Recruitment and Sample Acquisition

CRC, Breast cancer, and GBM samples were obtained from the University Health Network BioBank; AML samples were obtained from the University Health Network Leukemia BioBank; Lastly, healthy controls were recruited through the Family Medicine Centre at Mount Sinai Hospital (MSH) in Toronto, Canada. All samples

collected with patient consent, were obtained with institutional approval from the Research Ethics Board, from University Health Network and Mount Sinai Hospital in Toronto, Canada.

Specimen Processing – cfDNA

EDTA and ACD plasma samples were obtained from the BioBanks and from the Family Medicine Centre at Mount Sinai Hospital (MSH) in Toronto, Canada. All samples were either stored at -80˚C or in vapour phase liquid nitrogen until use. Cell-free DNA was extracted from 0.5-3.5 ml of plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen). The extracted DNA was quantified through Qubit prior to use.

Specimen Processing – PDX cfDNA

Human colorectal tumor tissue obtained with patient consent from the University Health Network Biobank as approved by the Research Ethics Board at University Health Network, was digested to single cells using collagenase A. Single cells were subcutaneously injected into 4-6 week old NOD/SCID male mouse. Mice were euthanized by CO2 inhalation prior to blood collection by cardiac puncture and stored in EDTA tubes. From the collected blood samples, the plasma was isolated and stored at -80C. Cell-free DNA was extracted from 0.3-0.7 ml of plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen). All animal work was carried out in compliance with the ethical regulations approved by the Animal Care Committee at University Health Network.

cfMeDIP-seq

A schematic representation of the cfMeDIP-seq protocol is shown in WO2017/190215. Prior to cfMeDIP, the DNA samples were subjected to library preparation using the Kapa Hyper Prep Kit (Kapa Biosystems). The manufacturer protocol was followed with some modifications. Briefly, the DNA of interest was added to 0.2 mL PCR tube and subjected to end-repair and A-Tailing. Adapter ligation was followed using NEBNext adapter (from the NEBNext Multiplex Oligos for Illumina kit, New England Biolabs) at a final concentration of 0.181 μM, incubated at 20˚C for 20 mins and purified with AMPure XP beads. The eluted library was digested using the USER enzyme (New

England Biolabs Canada) followed by purification with Qiagen MinElute PCR Purification Kit prior to MeDIP.

The prepared libraries were combined with the pooled methylated/unmethylated λ PCR product to a final DNA amount of 100 ng and subjected to MeDIP using the protocol from Taiwo et al. 2012[7] with some modifications. Briefly, for MeDIP, the Diagenode MagMeDIP kit (Cat# C02010021) was used following the manufacturer's protocol with some modifications. After the addition of 0.3 ng of the control methylated and 0.3ng of the control unmethylated *A. thaliana* DNA, the filler DNA (to complete the total amount of DNA [cfDNA + Filler + Controls] to 100ng) and the buffers to the PCR tubes containing the adapter ligated DNA, the samples were heated to 95˚C for 10 mins, then immediately placed into an ice water bath for 10 mins. Each sample was partitioned into two 0.2 mL PCR tubes: one for the 10% input control and the other one for the sample to be subjected to immunoprecipitation. The included 5-mC monoclonal antibody 33D3 (Cat#C15200081) from the MagMeDIP kit was diluted 1:15 prior to generating the diluted antibody mix and added to the sample. Washed magnetic beads (following manufacturer instructions) were also added prior to incubation at 4˚C for 17 hours. The samples were purified using the Diagenode iPure Kit and eluted in 50 µl of Buffer C. The success of the reaction (QC1) was validated through qPCR to detect the presence of the spiked-in *A. thaliana* DNA, ensuring a % recovery of unmethylated spiked-in DNA <1% and the % specificity of the reaction >99% (as calculated by 1-[recovery of spiked-in unmethylated control DNA over recovery of spiked-in methylated control DNA]), prior to proceeding to the next step. The optimal number of cycles to amplify each library was determined through the use of qPCR, after which the samples were amplified using the KAPA HiFi Hotstart Mastermix and the NEBNext multiplex oligos added to a final concentration of 0.3 µM. The PCR settings used to amplify the libraries were as follows: activation at 95˚C for 3 min, followed by predetermined cycles of 98˚C for 20 sec, 65˚C for 15 sec and 72˚C for 30 sec and a final extension of 72˚C for 1 min. The amplified libraries were purified using MinElute PCR purification column and then gel size selected with 3% Nusieve GTG agarose gel to remove any adapter dimers. Prior to submission for sequencing, the fold enrichment of a methylated human DNA region (testis-specific H2B, TSH2B) and an unmethylated human DNA region (GAPDH promoter) was determined for the MeDIP-seq and cfMeDIP-seq libraries generated from the HCT116 cell line DNA sheared to mimic cell free DNA (Cell line obtained from ATCC, mycoplasma free). The final libraries were

submitted for BioAnalyzer analysis prior to sequencing at the UHN Princess Margaret Genomic Centre on an Illumina HiSeq 2000.

Ultra-deep targeted sequencing for point mutation detection

We used the QIAgen Circulating Nucleic Acid kit to isolate cell-free DNA from ~20 mL of plasma (4-5x 10mL EDTA blood tubes) from patients with matched tumor tissue molecular profiling data generated prior to enrolment in early phase clinical trials at the Princess Margaret Cancer Centre. DNA was extracted from cell lines (dilution of CRC and MM cell lines) using the PureGene Gentra kit, fragmented to ~180 bp using a Covaris sonicator, and larger size fragments excluded using Ampure beads to mimic the fragment size of cell-free DNA. DNA sequencing libraries were constructed from 83 ng of fragmented DNA using the KAPA Hyper Prep Kit (Kapa Biosystems, Wilmington, MA) utilizing NEXTflex-96 DNA Barcode adapters (Bio Scientific, Austin, TX) adapters. To isolate DNA fragments containing known mutations, we designed biotinylated DNA capture probes (xGen Lockdown Custom Probes Mini Pool, Integrated DNA Technologies, Coralville, IA) targeting mutation hotspots from 48 genes tested by the clinical laboratory using the Illumina TruSeq Amplicon Cancer Panel. The barcoded libraries were pooled and then applied the custom hybrid capture library following manufacturer's instructions (IDT xGEN Lockdown protocol version 2.1). These fragments were sequenced to >10,000X read coverage using an Illumina HiSeq 2000 instrument. Resulting reads were aligned using bwa-mem and mutations detected using samtools and muTect version 1.1.4.

Modelling relationships between number of tumor-specific features and probability of detection by sequencing depth

We created 145,000 simulated genomes, with the proportion of cancer-specific methylated DMRs set to 0.001%, 0.01%, 0.1%, 1%, and 10% and consisting of 1, 10, 100, 1000 and 10000 independent DMRs respectively. We sampled 14,500 diploid genomes (representing 100ng of DNA) from these original mixtures and further sampled 10, 100, 1000, and 10000 reads per locus to represent sequencing coverage at those depths. This process was repeated 100 times for each combination of coverage, abundance, and number of features. We estimated the frequency of successful detection of at least 1 DMR for each combination of parameters and plotted probability curves (Figure 1A) to visually evaluate the influence of the number of features on the probability of successful detection conditional on sequencing depths.

Derivation of tissue-distinctive features, development of a multi-tissue classifier and validation in 450k data

cfDNA MeDIP profiles were quantified using the MEDIPS R package[8], converted to RPKMs, and afterwards transformed into log2 counts-per-million. Subsequently, a linear model was fit using limma-trend[9] on a matrix of features that mapped to FANTOM5 enhancers, CpG Islands, CpG shores and CpG Shelves, with the percentage of spike-in methylated DNA recovered included as a covariate to control for pulldown efficiency variation. Pairwise contrasts were evaluated for each pair of tissue types and the top 150 and the bottom 150 DMRs were selected for elastic net classifier training and validation of cancer-type specificity. Performance metrics were derived by majority class votes on out-of-fold calls from the model with the highest Kappa value in cross-validation, a heuristic previously employed in Chakravarthy et al[10].

Machine learning analyses for evaluation of classification accuracy

Model training and evaluation on the discovery cohort

In order to evaluate the performance of cfMeDIP data in tumor classification without high computational cost, we reduced the initial set of possible candidate features to windows encompassing CpG Islands, shores, shelves and FANTOM5 enhancers (hereby labelled "regulatory features"), yielding a matrix of 196 samples and 505,027 features. We then used the caret R package to partition the discovery cohort data into 50 independent training and test sets in an 80%-20% manner (Figure 2A). The splits were performed while class proportions across the discovery cohort were maintained. Then, we selected the top 300 DMRs by moderated t-statistic (150 hypermethylated, 150 hypomethylated) on the training data partition using limma-trend for each class versus other classes. A binomial GLMnet was then trained using these DMRs (up to 300 DMRs x 7 other classes = 2100 features) with the use of 3 iterations of 10-Fold Cross-Validation (CV) to optimize values of the mixing parameter (alpha, values = 0, 0.2, 0.5, 0.8 and 1) and the penalty (lambda, values = 0 – 0.05 in increments of 0.01) using Cohen's Kappa as the performance metric. For each training set, this yielded a collection of 6 one-class vs-other-classes binomial classifiers.

We then estimated classification performance on the held-out test set using the AUROC (area under the receiver operating characteristic curve). These estimates

represent unbiased measures of classification, as the held-out test set samples were not used for either DMR pre-selection or GLMnet training and tuning. The 50 independent training and test sets also permitted for minimization of optimistic estimates due to training-set bias.

Model evaluation on the validation cohort

For each validation cohort cfMeDIP sample, we estimated class probabilities for the AML, LUC and normal one-vs-all binomial classifiers trained on the 50 different training sets within the discovery cohort. The probabilities from the 50 models were averaged to produce a single score that was then used for AUROC estimation. We also evaluated if disease stage affected performance by estimating AUROC when either early (Stages I and II) or late stage LUC samples (Stages III and IV) were left out for the one-vs-all classifier.

Results and Discussion

We bioinformatically simulated mixtures with different proportions of ctDNA, from 0.001% to 10% (Figure 1A, column facets). We also simulated scenarios where the ctDNA had 1, 10, 100, 1000, or 10000 DMRs (Differentially Methylated Regions) as compared to normal cfDNA (Figure 1A, row facets). Reads were then sampled at varying sequencing depths at each locus (10X, 100X, 1000X, and 10000X) (Figure 1A, x-axis). We found an increasing probability of detecting of at least 1 cancer-specific event (Figure 1A) as the number of DMRs increased, even at low abundance of cancer ctDNA and shallow coverage.

Moreover, pan-cancer data from The Cancer Genome Atlas (TCGA) shows large numbers of DMRs between tumor and normal tissues across virtually all tumor types[11]. Therefore, these findings highlighted that an assay that successfully recovered cancer-specific DNA methylation alterations from ctDNA could serve as a very sensitive tool to detect, classify, and monitor malignant disease with low sequencing-associated costs.

However, genome-wide mapping of DNA methylation in plasma cfDNA is challenging due to the very low quantities and fragmentation of DNA in circulation[12]. As a result, previous efforts at methylation profiling of cfDNA has mainly been restricted to locus specific PCR-based assays[2, 3], such as an FDA approved *SEPT9* methylation assay

for colorectal cancer screening[13]. While recent efforts have been made to perform whole-genome bisulfite-sequencing of fragmented cfDNA[14-16], the low genome-wide abundance of CpGs is likely to reduce the amount of useful methylation-related information available from sequencing. Therefore, the main issues with WGBS on plasma DNA are the high cost, low efficiency, and DNA losses associated with the bisulfite conversion. On the other hand, a method that selectively enriches for CpG-rich features prone to methylation is likely to maximize the amount of useful information available per read, decrease the cost, and decrease the DNA losses.

A genome-wide method suitable for cfDNA methylation mapping

We developed a new method termed cfMeDIP-seq (cell-free Methylated DNA Immunoprecipitation and high-throughput sequencing) to perform genome-wide DNA methylation mapping using cell-free DNA. The cfMeDIP-seq method described here was developed through the modification of an existing low input MeDIP-seq protocol[7] that in our experience is very robust down to 100 ng of input DNA. However, the majority of plasma samples yield much less than 100 ng of DNA. To overcome this challenge, we added exogenous λ DNA (filler DNA) to the adapter-ligated cfDNA library in order to artificially inflate the amount of starting DNA to 100 ng. This minimizes the amount of non-specific binding by the antibody and also minimizes the amount of DNA lost due to binding to plasticware. The filler DNA consisted of amplicons similar in size to an adapter-ligated cfDNA library and was composed of unmethylated and *in vitro* methylated DNA at different CpG densities. The addition of this filler DNA also serves a practical use, as different patients will yield different amounts of cfDNA, allowing for the normalization of input DNA amount to 100 ng. This ensures that the downstream protocol remains exactly the same for all samples regardless of the amount of available cfDNA.

We first validated the cfMeDIP-seq protocol using DNA from human colorectal cancer cell line HCT116, sheared to a fragment size similar to that observed in cfDNA. HCT116 was chosen because of the availability of public DNA methylation data. We simultaneously performed the gold standard MeDIP-seq protocol[7] using 100 ng of sheared cell line DNA and the cfMeDIP-seq protocol using 10 ng, 5 ng, and 1 ng of the same sheared cell line DNA. This was performed in two biological replicates. For all the conditions, we obtained more than 99% specificity of the reaction (1- [recovery of spiked-in unmethylated control DNA over recovery of spiked-in methylated control

DNA]), and a very high enrichment of a known methylated region over an unmethylated region (TSH2B0 and GAPDH, respectively) (Figure 6B).

The libraries were sequenced to saturation (Figure 6A) at around 30 to 70 million reads per library (Supplementary Table 1). The raw reads were aligned to both the human genome and the λ genome, and found virtually no alignment was found to the λ genome (Supplementary Table 1). Therefore, the addition of the exogenous λ DNA as filler DNA did not interfere with the generation of sequencing data. Finally, we calculate the CpG Enrichment Score as a quality control measure for the immunoprecipitation step[8]. All the libraries showed similar enrichment for CpGs while the input control, as expected, showed no enrichment (Figure 6C), validating our immunoprecipitations even at extremely low inputs (1ng).

Genome-wide correlation estimates comparing different input DNA levels show that both MeDIP-seq (100 ng) and cfMeDIP-seq (10, 5, and 1 ng) methods were very robust, with Pearson correlation of at least 0.94 between any two biological replicates (Figure 1B). The analysis also demonstrates that cfMeDIP-seq at 5 and 10 ng of input DNA can robustly recapitulate the methylation profile obtained by traditional MeDIP-seq at 100 ng (Pairwise Pearson correlation of at least 0.9) (Figure 1B). The performance of cfMeDIP-seq at 1 ng of input DNA is reduced compared to MeDIP-seq at 100 ng but still shows a strong Pearson correlation at >0.7 (Figure 1B). We also observed that the cfMeDIP-seq protocol recapitulates the DNA methylation profile of HCT116 using gold standard RRBS (Reduced Representation Bisulfite Sequencing) and WGBS (Whole-Genome Bisulfite Sequencing) (Figure 1C). Altogether, our data suggests that cfMeDIP-seq is a robust protocol for genome-wide methylation mapping of fragmented and low input DNA material, such as circulating cfDNA.

cfMeDIP-seq displays high-sensitivity for detection of tumor-derived ctDNA

To evaluate the sensitivity of the cfMeDIP-seq protocol, we performed a serial dilution of Colorectal Cancer (CRC) HCT116 cell line DNA into a Multiple Myeloma (MM) MM1.S cell line DNA, both sheared to mimic cfDNA sizes. We diluted the CRC DNA from 100%, 10%, 1%, 0.1%, 0.01%, 0.001%, to 0% and performed cfMeDIP-seq on each of these dilutions. We also performed ultra-deep (10,000X median coverage) targeted sequencing for detection of three point mutations in the same samples. The observed number of DMRs identified at each CRC dilution point versus the pure MM

DNA using a 5% False Discovery rate (FDR) threshold was almost perfectly linear (r^2=0.99, p<0.0001) with the expected number of DMRs based on the dilution factor (Figure 1D) down to a 0.001% dilution. Moreover, the DNA methylation signal within these DMRs also shows almost perfect linearity (r^2=0.99, p<0.0001) between the observed versus expected signal (Figure 1E; Supplementary Table 2B). In comparison, beyond the 1% dilution, ultra-deep targeted sequencing could not reliably distinguish between the CRC-specific variants and the spurious variants due to PCR or sequencing-errors (Figure 1F; Supplementary Table 2A). Thus, cfMeDIP-seq displays excellent sensitivity for the detection of cancer-derived DNA, exceeding the performance of variant detection by ultra-deep targeted sequencing using a standard protocol.

Cancer DNA is frequently hypermethylated at CpG-rich regions[17]. Since cfMeDIP-seq specifically targets methylated CpG-rich sequences, we hypothesized that ctDNA would be preferentially enriched during the immunoprecipitation procedure. To test this, we generated patient-derived xenografts (PDXs) from two colorectal cancer patients and collected the mouse plasma. Tumor-derived human cfDNA was present at less than 1% frequency within the total cfDNA pool in the input samples and at 2-fold greater abundance following immunoprecipitation (Figure 1G; Supplementary Table 3). These results suggest that through biased sequencing of ctDNA, the cfMeDIP procedure could further increase ctDNA detection sensitivity.

Circulating plasma cfDNA methylation profile can distinguish between multiple cancer types and healthy donors

DNA methylation patterns are tissue-specific, and have been used to stratify cancer patients into clinically relevant disease subgroups in glioblastoma[18], ependymomas[6], colorectal[19], and breast[20, 21], among many other cancer types. We asked if cfDNA associated profiles could be used to identify tissues-of-origin for multiple tumor types. To this end, we profiled 196 samples from 5 different tumor types and normal controls from early and late stage tumors. We used linear modeling to identify the top 300 DMRs mapping to CpG shores, shelves, islands and FANTOM5 enhancers for each pairwise comparison, leading to a total of 2,100 unique DMRs (Figure 2A). Density clustering based on t-Distributed Stochastic Neighbor Embedding (tSNE)[22] of the 196 plasma samples based on the methylation status of these features revealed distinct clustering of samples based on tissue-of-origin and tumor

types (Figure 2B,C). Using an elastic net multi-cancer classifier fit with these features (Figure 2A), we observed highly accurate discrimination between different tumor types (Figure 2D).

Discrimination of disease subtypes

We evaluated the ability of cfDNA MeDIP profiles to discriminate between disease subtypes in five distinct cases – gene expression pattern (ER status in breast cancer), copy number aberration (HER2 status in breast cancer), rearrangement (FLT3 ITD status in AML), point mutation (IDH mutation in GBM), and finally histology in lung cancer. In each case, linear models were used to select and rank features as described earlier. In each case, hierarchical clustering was used to evaluate the grouping of samples. Density clustering based on t-Distributed Stochastic Neighbor Embedding (tSNE)[22] based on the methylation status of selected features revealed distinct clustering of samples based on each of these five distinct examples of cancer subtype classification.

Detection of cancers and classification of cancer types using machine learning

In order to rigorously evaluate the ability of cfMeDIP profiles to detect cancers and further classify cancer types, we then conducted a set of machine learning analyses on our discovery cohort. To allow for accelerated computational analysis, we initially reduced our cfMeDIP discovery cohort to features mapping to CpG islands, shores, shelves and FANTOM5 enhancers (n=505,027 windows). We then implemented a strategy on our discovery cohort samples to derive unbiased estimates of performance, while accounting for training-set biases.

Herein, we split the discovery cohort into balanced training and test sets (80% training set, 20% test set). Using only the samples in the training set, we selected the top 300 DMRs for each class (sample type) versus other classes, based on limma-trend test statistics, and trained a series of one-versus-other-classes GLMnets using these features on the training set data. The training procedure consisted of 3 rounds of 10-Fold Cross-Validation (CV) across a grid of values for alpha and lambda with optimisation for Cohen's Kappa. The use of multiple rounds of 10-Fold CV was motivated by a desire to leverage additional randomisation for more generalisable model tuning.

Performance was then evaluated using AUROC (area under the receiver operating characteristic curve) derived from test set samples (held-out during the DMR selection and the subsequent GLMnet training/tuning steps). This process was repeated with 50 different splits of the discovery cohort into training and test sets to mitigate the influence of training-set biases. This culminated in a collection of 50 models for each one-vs other-classes comparison (480 models in total). Hereby, we refer to this collection of models as E50.

Subsequently, we evaluated performance across batches by generating a validation cohort of additional 152 plasma samples: AML (n=35), lung cancer (n=55) and healthy control (n=62) samples. For each class, we averaged the class probabilities output by the models in E50, and estimated AUROC for the one class vs. all others classes (Figure 3A). The classifiers showed high AUROC values for the classification of AML vs others (0.993), LUC vs others (0.943) and normal vs others (1.000). This further confirmed the ability of cfMeDIP-seq coupled with a machine learning approach to accurately detect and classify tumor type. Finally, we observed that the classifiers were as accurate in early stage samples (0.950) as in late stage samples (0.934) (Figure 3B), suggested that this approach is applicable for cancer early detection and for detection of cancer at both early stages and late stages.

Additional advantages of cfDNA methylome profiling with cfMeDIP-seq

The ability of cfDNA methylation patterns to accurately represent tissue-of-origin also overcomes limitations of mutation-based assays, wherein specificity for tissues-of-origin may be low due to the recurrent nature of many potential driver mutations across cancers in different tissues[23]. Mutation based assays may also be rendered insensitive by the clonal structure of tumors, where subclonal drivers may be harder to detect by virtue of lower abundance in ctDNA[24]. Mutation based ctDNA approaches are also vulnerable to potential confounding by driver mutations in benign tissues, which have been observed[25], and documented to display evidence of positive selection[26].

Taken together, our findings - based on the largest collection of cancer cfDNA methylomes derived to date - establish cfMeDIP-seq as an efficient and cost-effective tool with the potential to influence management of cancer and early detection. The accuracy and versatility of cfMeDIP-seq may be useful to inform therapeutic decisions in settings where resistance is correlated to epigenetic alterations, such as sensitivity

to androgen receptor inhibition in prostate cancer[27]. The potential opportunities for early diagnosis and screening may be particularly evident in lung cancer, a disease in which screening has already shown clinical utility but for which existing screening tests (i.e., low dose CT scanning) has significant limitations such as ionizing radiation exposure and high false positive rate.

In conclusion, our findings underscore the utility of cfDNA methylation profiles as a basis for non-invasive, cost-effective, sensitive, highly accurate early tumor detection, multi-cancer classification, and cancer subtype classification.

Although preferred embodiments of the invention have been described herein, it will be understood by those skilled in the art that variations may be made thereto without departing from the spirit of the invention or the scope of the appended claims. All documents disclosed herein, including those in the following reference list, are incorporated by reference.

Table 1: Number of reads and mapping efficiency of sequenced MeDIP-seq (100 ng Rep 1 and Rep 2) and cfMeDIP-seq (10 ng, 5 ng and 1 ng, Rep1 and Rep 2) libraries prepared using various starting inputs of HCT116 cell line DNA sheared to mimic cfDNA, to human (Hg19) genome and λ genome. Two biological replicates were use for starting input DNA. For starting inputs less than 100 ng, the samples were topped up with exogenous λ DNA to artificially increase the starting amount to 100 ng prior to MeDIP.

Sample	# of raw reads	# of aligned reads to human genome (Hg19)	Mapping efficiency to human genome (Hg19)	# of aligned reads to λ genome	Mapping efficiency to λ genome
Input	74,504,053	71,343,168	95.76	12	0.00
100 ng Replicate 1	55,396,238	50,472,273	91.11	0	0.00
100 ng Replicate 2	66,569,209	60,770,277	91.29	1	0.00
10 ng Replicate 1	70,054,607	64,020,441	91.39	0	0.00
10 ng Replicate 2	58,297,539	53,308,777	91.44	0	0.00
5 ng Replicate 1	65,845,430	60,540,743	91.94	1	0.00
5 ng Replicate 2	64,750,879	59,358,412	91.67	0	0.00
1 ng Replicate 1	35,102,361	32,258,451	91.90	0	0.00
1 ng Replicate 2	33,881,118	31,194,711	92.07	0	0.00

Table 2A: Mean coverage of ultra-deep targeted variant sequencing using dilution series of CRC cell line HCT116 DNA into MM cell line MM1.S DNA

Dilution (% of CRC DNA)	Uncollapsed reads mean target coverage	SSCS (single strand consensus sequences) mean target coverage	DCS (duplex consensus sequences) mean target coverage
100	155,964	4284	655
10	154,657	4877	654
1	154,419	4890	654
0.1	183,271	5674	887
0.01	238,291	8068	1602
0.001	199,766	7337	1299
0.0001	187,695	6891	1181
0	216,434	7721	1412

Table 2B: Resultant observed DMRs and DNA methylation signal from the dilution series of CRC cell line HCT116 DNA into MM cell line MM1.S DNA

Dilution (% of CRC DNA)	Observed number of DMRs	Observed DNA methylation signal (sum of RPKMs within DMRs)
100	111,472	645,683.90
10	1,597	8,775.61
1	692	4,521.60
0.1	12	75.71
0.01	8	79.73
0.001	2	22.42

Table 3: Number of reads and mapping efficiency of cfMeDIP-seq libraries of PDX and Input Control samples after aligning to human (Hg19) genome

Sample	# of Raw reads	# of Aligned reads to human genome (Hg19)	Mapping efficiency to human genome (Hg19)
Input Control 1	46,857,633	389,073	0.83
Input Control 2	35,658,454	283,799	0.80
PDX 1	49,997,949	1,080,277	2.16
PDX 2	34,802,767	614,988	1.77

5

Reference List

1. Diaz, L.A., Jr. and A. Bardelli, *Liquid biopsies: genotyping circulating tumor DNA.* J Clin Oncol, 2014. **32**(6): p. 579-86.
2. Lehmann-Werman, R., et al., *Identification of tissue-specific cell death using methylation patterns of circulating DNA.* Proc Natl Acad Sci U S A, 2016. **113**(13): p. E1826-34.
3. Visvanathan, K., et al., *Monitoring of Serum DNA Methylation as an Early Independent Marker of Response and Survival in Metastatic Breast Cancer: TBCRC 005 Prospective Biomarker Study.* J Clin Oncol, 2016: p. JCO2015662080.
4. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nat Med, 2014. **20**(5): p. 548-54.
5. Aravanis, A.M., M. Lee, and R.D. Klausner, *Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection.* Cell, 2017. **168**(4): p. 571-574.
6. Mack, S.C., et al., *Epigenomic alterations define lethal CIMP-positive ependymomas of infancy.* Nature, 2014. **506**(7489): p. 445-50.
7. Taiwo, O., et al., *Methylome analysis using MeDIP-seq with low DNA concentrations.* Nat Protoc, 2012. **7**(4): p. 617-36.
8. Lienhard, M., et al., *MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments.* Bioinformatics, 2014. **30**(2): p. 284-6.
9. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
10. Chakravarthy, A., et al., *Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx.* J Clin Oncol, 2016. **34**(34): p. 4132-4141.
11. Hoadley, K.A., et al., *Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.* Cell, 2014. **158**(4): p. 929-44.
12. Fleischhacker, M. and B. Schmidt, *Circulating nucleic acids (CNAs) and cancer--a survey.* Biochim Biophys Acta, 2007. **1775**(1): p. 181-232.
13. Potter, N.T., et al., *Validation of a real-time PCR-based qualitative assay for the detection of methylated SEPT9 DNA in human plasma.* Clin Chem, 2014. **60**(9): p. 1183-91.
14. Legendre, C., et al., *Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer.* Clin Epigenetics, 2015. **7**: p. 100.
15. Sun, K., et al., *Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments.* Proc Natl Acad Sci U S A, 2015. **112**(40): p. E5503-12.
16. Chan, K.C., et al., *Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing.* Proc Natl Acad Sci U S A, 2013. **110**(47): p. 18761-8.
17. Sharma, S., T.K. Kelly, and P.A. Jones, *Epigenetics in cancer.* Carcinogenesis, 2010. **31**(1): p. 27-36.
18. Sturm, D., et al., *Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma.* Cancer Cell, 2012. **22**(4): p. 425-37.

19. Hinoue, T., et al., *Genome-scale analysis of aberrant DNA methylation in colorectal cancer.* Genome Res, 2012. **22**(2): p. 271-82.
20. Stirzaker, C., et al., *Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value.* Nat Commun, 2015. **6**: p. 5899.
21. Fang, F., et al., *Breast cancer methylomes establish an epigenomic foundation for metastasis.* Sci Transl Med, 2011. **3**(75): p. 75ra25.
22. Laurens van der Maaten, G.H., *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008. **9**: p. 2579-2605.
23. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-9.
24. McGranahan, N., et al., *Clonal status of actionable driver events and the timing of mutational processes in cancer evolution.* Sci Transl Med, 2015. **7**(283): p. 283ra54.
25. Zauber, P., S. Marotta, and M. Sabbath-Solitare, *KRAS gene mutations are more common in colorectal villous adenomas and in situ carcinomas than in carcinomas.* Int J Mol Epidemiol Genet, 2013. **4**(1): p. 1-10.
26. Martincorena, I., et al., *Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin.* Science, 2015. **348**(6237): p. 880-6.
27. Beltran, H., et al., *Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer.* 2016. **22**(3): p. 298-305.

CLAIMS:

1. A method of detecting the presence of DNA from cancer cells in a subject comprising:

   (a) providing a sample of cell-free DNA from a subject;

   (b) subjecting the sample to library preparation to permit subsequent sequencing of the cell-free methylated DNA;

   (c) adding a first amount of filler DNA to the sample, wherein at least a portion of the filler DNA is methylated, then optionally denaturing the sample;

   (d) capturing cell-free methylated DNA using a binder selective for methylated polynucleotides;

   (e) sequencing the captured cell-free methylated DNA;

   (f) comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals;

   (g) identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals.

2. The method of claim 1, wherein the sample is from the subject's blood or plasma.

3. The method of claim 1 or 2, wherein comparison step (f) is based on fit using a statistical classifier.

4. The method of claim 3, wherein the classifier is machine learning-derived.

5. The method of claim 4, wherein the classifier is an elastic net classifier, lasso, support vector machine, random forest, or neural network.

6. The method of any one of claims 1-5, wherein the control cell-free methylated DNAs sequences from healthy and cancerous individuals are comprised in a database of Differentially Methylated Regions (DMRs) between healthy and cancerous individuals.

7. The method of any one of claims 1-6, wherein the control cell-free methylated DNA sequences from healthy and cancerous individuals are limited to those control cell-free methylated DNA sequences which are differentially methylated as between healthy and cancerous individuals in DNA derived from cell-free DNA.

8. The method of claim 7, wherein the control cell-free methylated DNA sequences are differentially methylated as between healthy and cancerous individuals in DNA derived from blood plasma.

9. The method of any one of claims 1-8, wherein the sample has less than 100 ng, 75 ng, or 50 ng of cell-free DNA.

10. The method of any one of claims 1-9, wherein the first amount of filler DNA comprises about 5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% methylated filler DNA with remainder being unmethylated filler DNA, and preferably between 5% and 50%, between 10%-40%, or between 15%-30% methylated filler DNA.

11. The method of any one of claims 1-9, wherein the first amount of filler DNA is from 20 ng to 100 ng, preferably 30 ng to 100 ng, more preferably 50 ng to 100 ng.

12. The method of any one of claims 1-11, wherein the cell-free DNA from the sample and the first amount of filler DNA together comprises at least 50 ng of total DNA, preferably at least 100 ng of total DNA.

13. The method of any one of claims 1-12, wherein the filler DNA is 50 bp to 800 bp long, preferably 100 bp to 600 bp long, and more preferably 200 bp to 600 bp long.

14. The method of any one of claims 1-13, wherein the filler DNA is double stranded.

15. The method of any one of claims 1-2, wherein the filler DNA is junk DNA.

16. The method of any one of claims 1-3, wherein the filler DNA is endogenous or exogenous DNA.

17. The method of claim 16, wherein the filler DNA is non-human DNA, preferably λ DNA.

18. The method of any one of claims 1-17, wherein the filler DNA has no alignment to human DNA.

19. The method of any one of claims 1-18, wherein the binder is a protein comprising a Methyl-CpG-binding domain.

20. The method of any one of claims 1-19, wherein the protein is a MBD2 protein.

21. The method of any one of claims 1-20, wherein step (d) comprises immunoprecipitating the cell-free methylated DNA using an antibody.

22. The method of claim 21, comprising adding at least 0.05 µg of the antibody to the sample for immunoprecipitation, and preferably at least 0.16 µg.

23. The method of claim 21, wherein the antibody is 5-MeC antibody.

24. The method of claim 21, further comprising the step of adding a second amount of control DNA to the sample after step (c) for confirming the immunoprecipitation reaction.

25. The method of any one of claims 1-23, further comprising the step of adding a second amount of control DNA to the sample after step (c) for confirming the capture of cell-free methylated DNA.

26. The method of any one of claims 1-25, wherein identifying the presence of DNA from cancer cells further includes identifying the cancer cell tissue of origin.

27. The method of claim 26, wherein identifying the cancer cell tissue of origin further includes identifying a cancer subtype.

28. The method of claim 27, wherein the cancer subtype differentiates the cancer based on stage, histology, gene expression pattern, copy number aberration, rearrangement, or point mutational status.

29. The method of any one of claims 1-28, wherein the comparison in step (f) is carried out genome-wide.

30. The method of any one of claims 1-28, wherein the comparison in step (f) is restricted from genome-wide to specific regulatory regions.

31. The method of claim 30, wherein the regulatory regions are FANTOM5 enhancers, CpG Islands, CpG shores, CpG Shelves, or any combination of the foregoing.

32. The method of any one of claims 1-31, wherein steps (f) and (g) are carried out by a computer processor.

33. A method of detecting the presence of DNA from cancer cells and identifying a cancer subtype, the method comprising:

a. receiving sequencing data of cell-free methylated DNA from a subject sample;

b. comparing the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals;

c. identifying the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the

captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals; and

d. if DNA from cancer cells from step c. is identified, further identifying the cancer cell tissue of origin and cancer subtype based on the comparison in step b.

34. A computer-implemented method of detecting the presence of DNA from cancer cells and identifying a cancer subtype, the method comprising:

a. receiving, at least one processor, sequencing data of cell-free methylated DNA from a subject sample;

b. comparing, at the at least one processor, the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals;

c. identifying, at the at least one processor, the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals and if DNA from cancer cells from step c. is identified, further identifying the cancer cell tissue of origin and cancer subtype based on the comparison in step b;

35. A computer program product for use in conjunction with a general-purpose computer having a processor and a memory connected to the processor, the computer program product comprising a computer readable storage medium having a computer mechanism encoded thereon, wherein the computer program mechanism may be loaded into the memory of the computer and cause the computer to carry out the method of claim 34.

36. A computer readable medium having stored thereon a data structure for storing the computer program product according to claim 35.

37. A device for detecting the presence of DNA from cancer cells and identifying a cancer subtype, the device comprising:

at least one processor; and

electronic memory in communication with the at one processor, the electronic memory storing processor-executable code that, when executed at the at least one processor, causes the at least one processor to:

a. receive sequencing data of cell-free methylated DNA from a subject sample;

b. compare the sequences of the captured cell-free methylated DNA to control cell-free methylated DNAs sequences from healthy and cancerous individuals;

c. identify the presence of DNA from cancer cells if there is a statistically significant similarity between one or more sequences of the captured cell-free methylated DNA and cell-free methylated DNAs sequences from cancerous individuals and if DNA from cancer cells from is identified, further identify the cancer cell tissue of origin and cancer subtype based on the comparison in step b.

38. A method of detecting the presence of DNA from cancer cells and determining the location of the cancer from which the cancer cells arose from two or more possible organs, the method comprising:

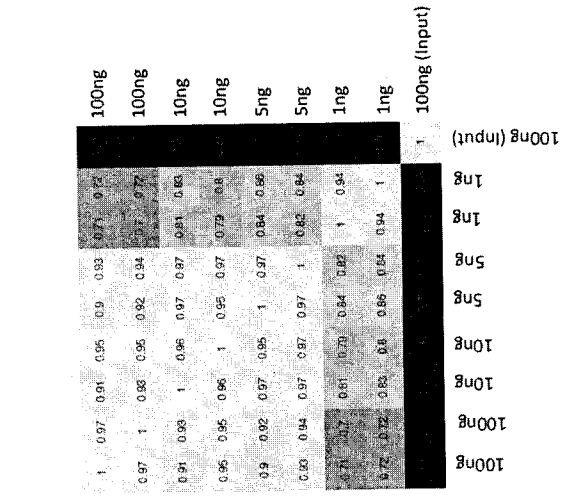
(a) providing a sample of cell-free DNA from a subject;

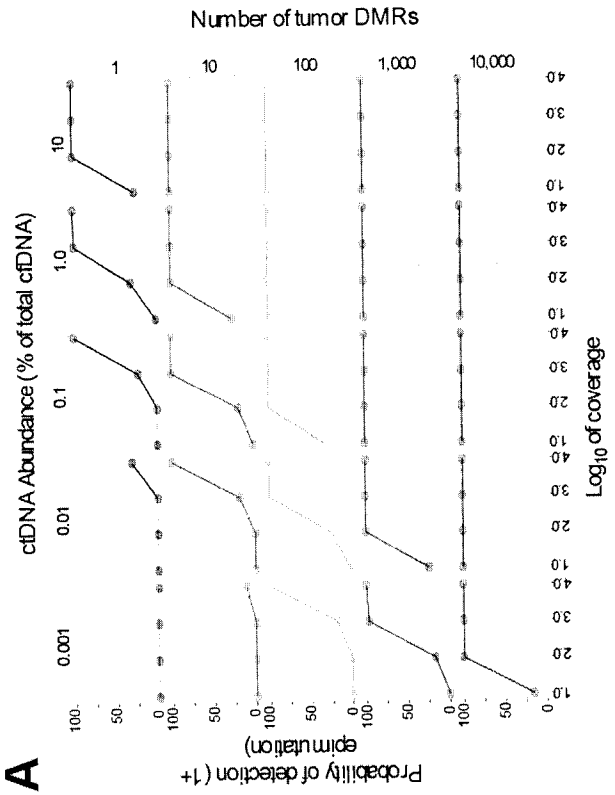
(b) capturing cell-free methylated DNA from said sample, using a binder selective for methylated polynucleotides;

(c) sequencing the captured cell-free methylated DNA;

(d) comparing the sequence patterns of the captured cell-free methylated DNA to DNAs sequence patterns of two or more population(s) of control individuals, each of said two or more populations having localized cancer in a different organ;

(e) determining as to which organ the cancer cells arose on the basis of a statistically significant similarity between the pattern of methylation of the cell-free DNA and one of said two or more populations.

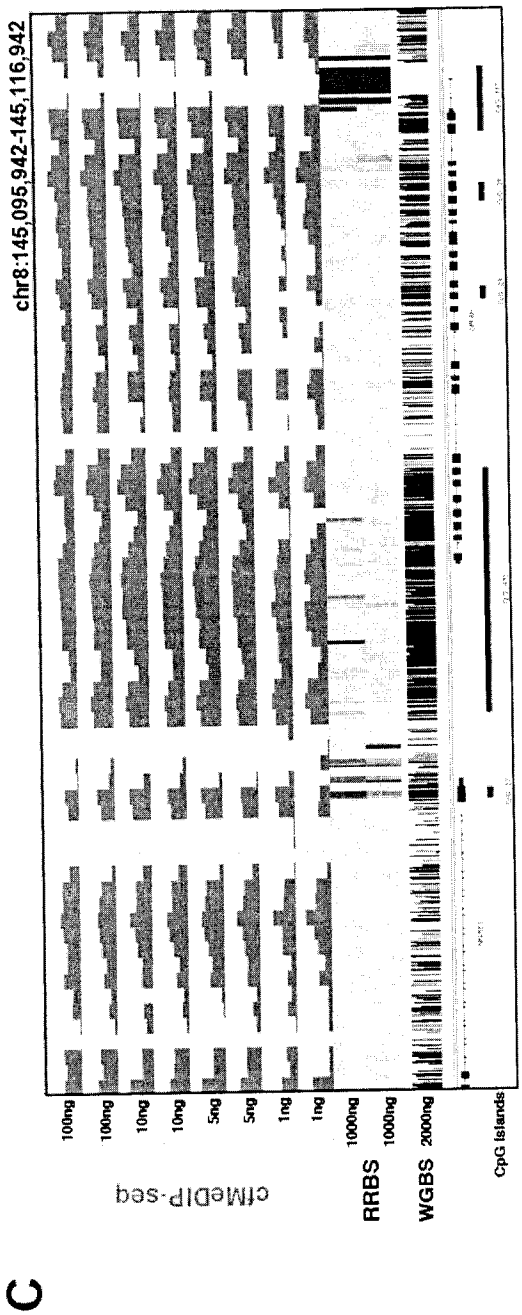




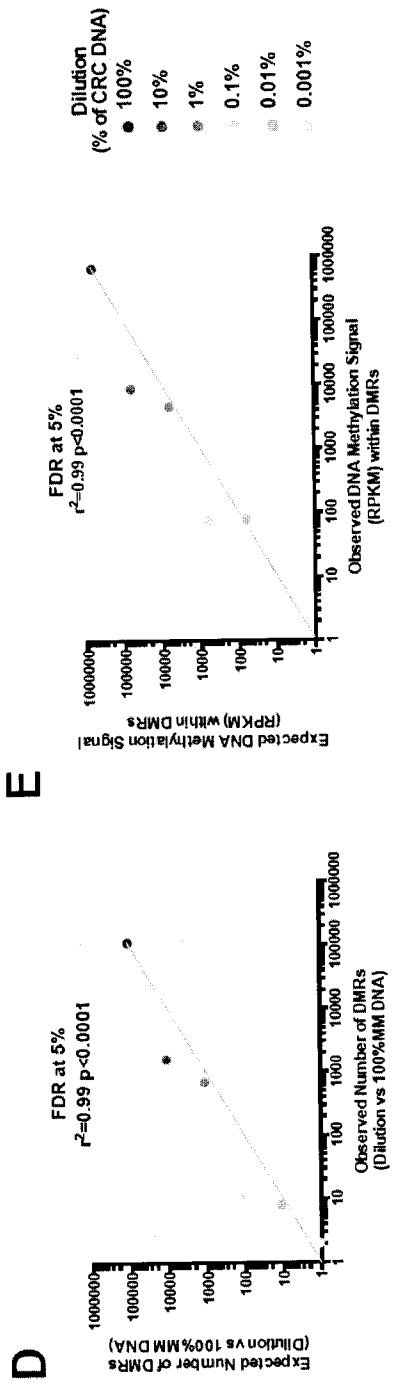
Figures 1A and 1B

Figure 1C

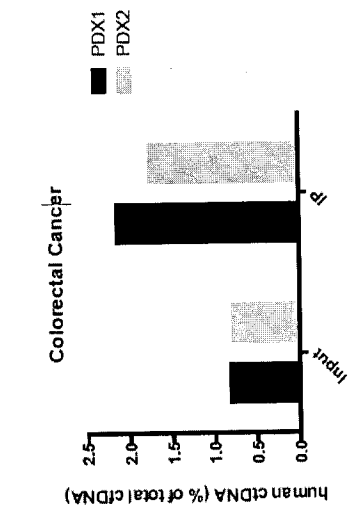
C


chr8:145,095,942-145,116,942
100ng
100ng
10ng
10ng
5ng
5ng
1ng
1ng
1000ng
1000ng
2000ng
cfMeDIP-seq
RRBS
WGBS
CpG Islands

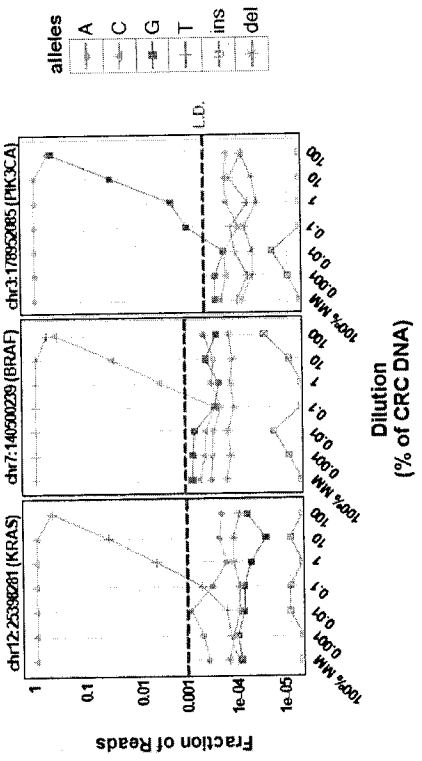
Figures 1D and 1E



Figures 1F and 1G

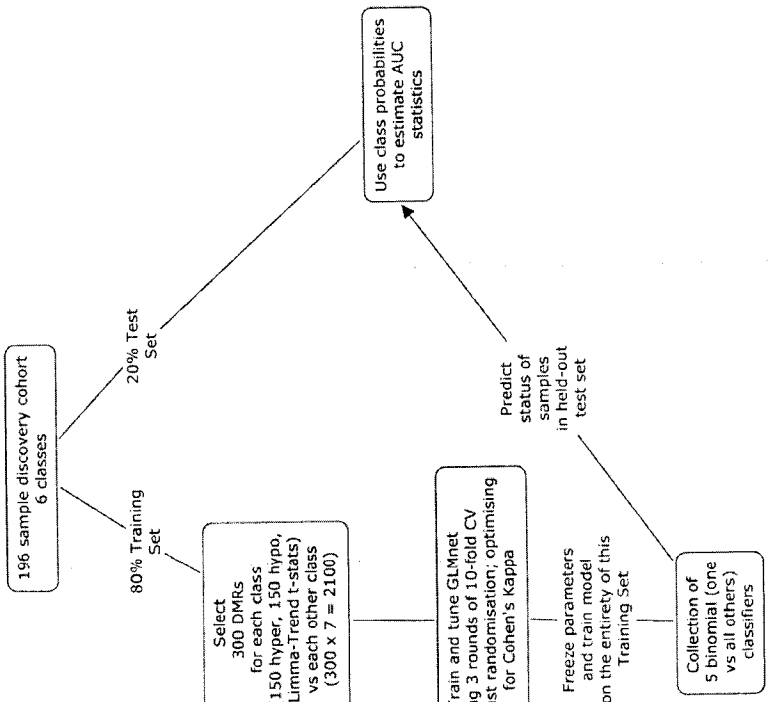


G



F

FIGURE 2A


196 sample discovery cohort
6 classes
20% Test Set
80% Training Set
Use class probabilities to estimate AUC statistics
Select 300 DMRs for each class (150 hyper, 150 hypo, Limma-Trend t-stats) vs each other class (300 x 7 = 2100)
Train and tune GLMnet using 3 rounds of 10-fold CV for robust randomisation; optimising for Cohen's kappa
Predict status of samples in held-out test set
Freeze parameters and train model on the entirety of this Training Set
Collection of 5 binomial (one vs all others) classifiers
Repeat process 50 times with different class-balanced splits of the discovery cohort into training & test sets.

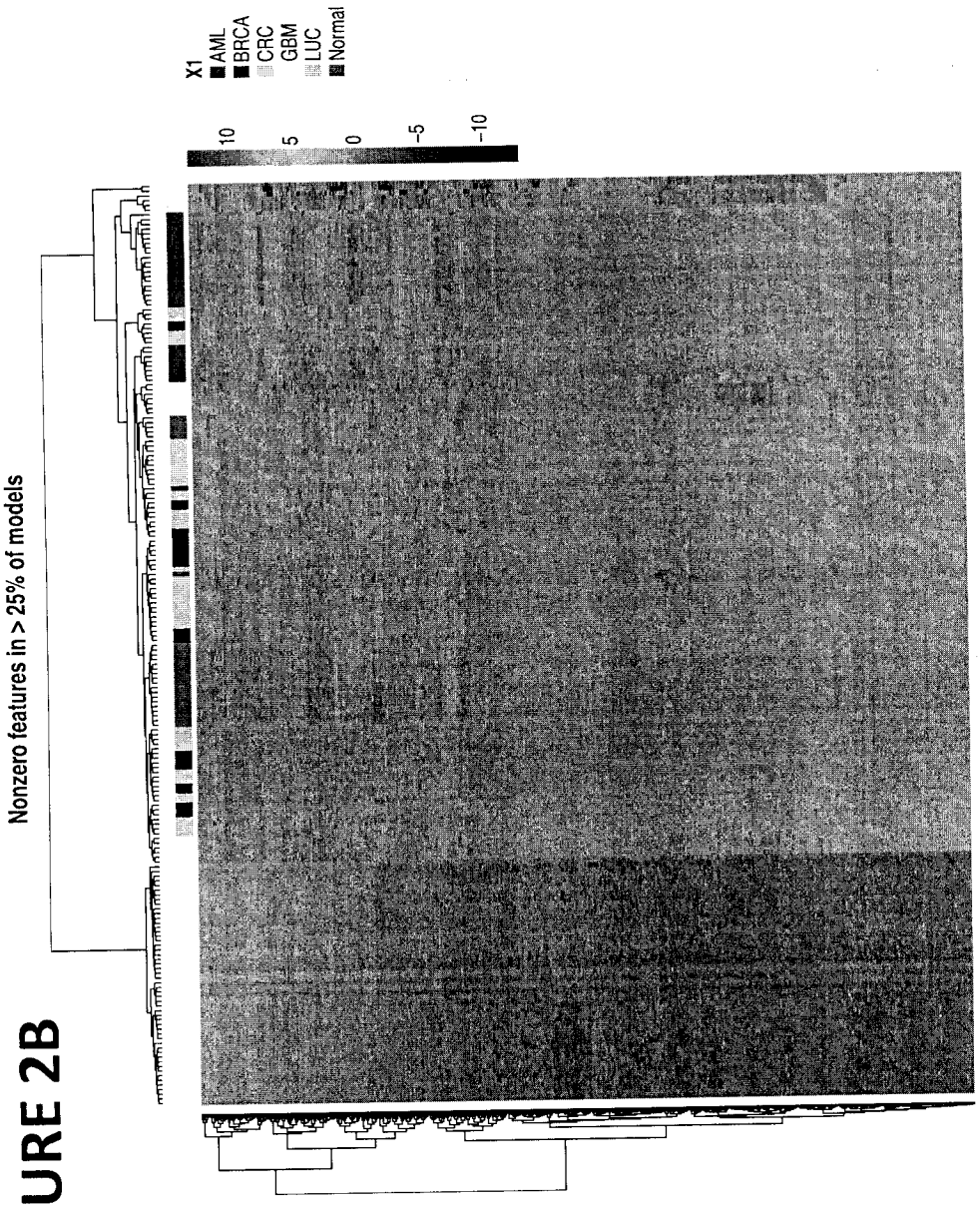
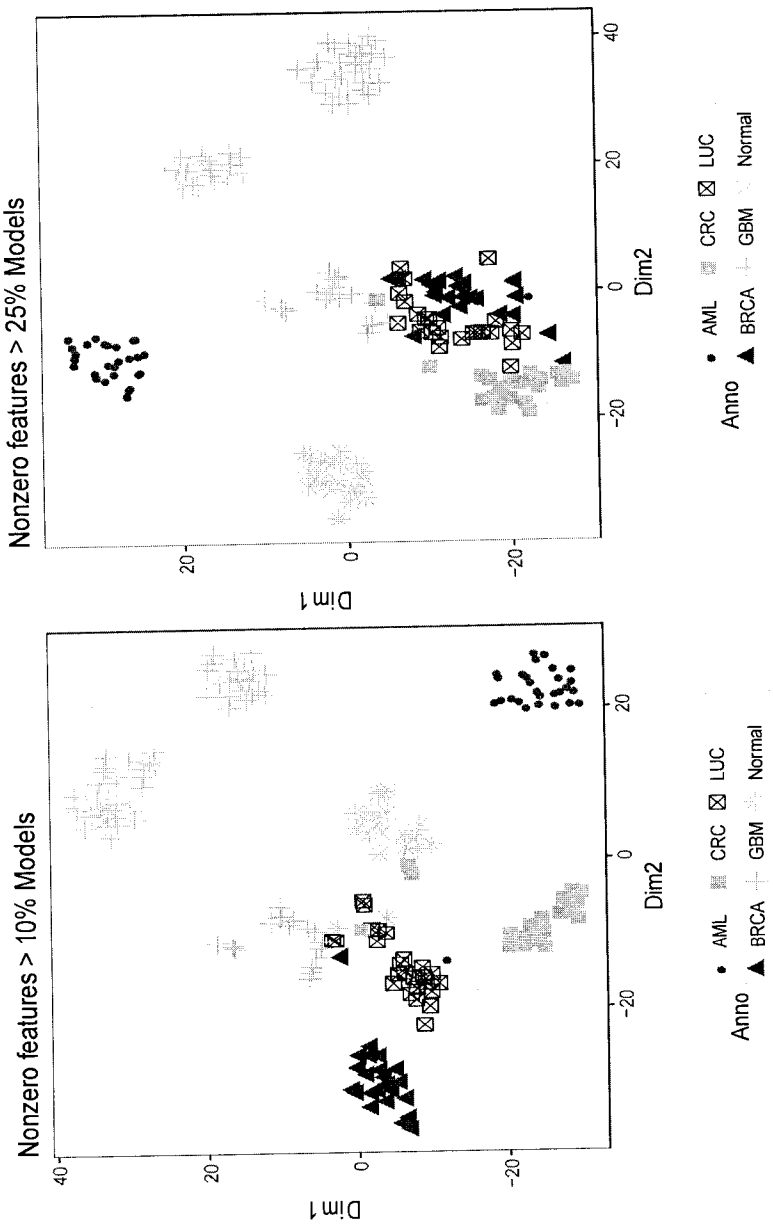

X1
AML
BRCA
CRC
GBM
LUC
Normal
10
5
0
−5
−10
Nonzero features in > 25% of models

FIGURE 2B

FIGURE 2C


Nonzero features > 25% Models
Nonzero features > 10% Models
Dim1
Dim2
Anno
AML
BRCA
CRC
GBM
LUC
Normal

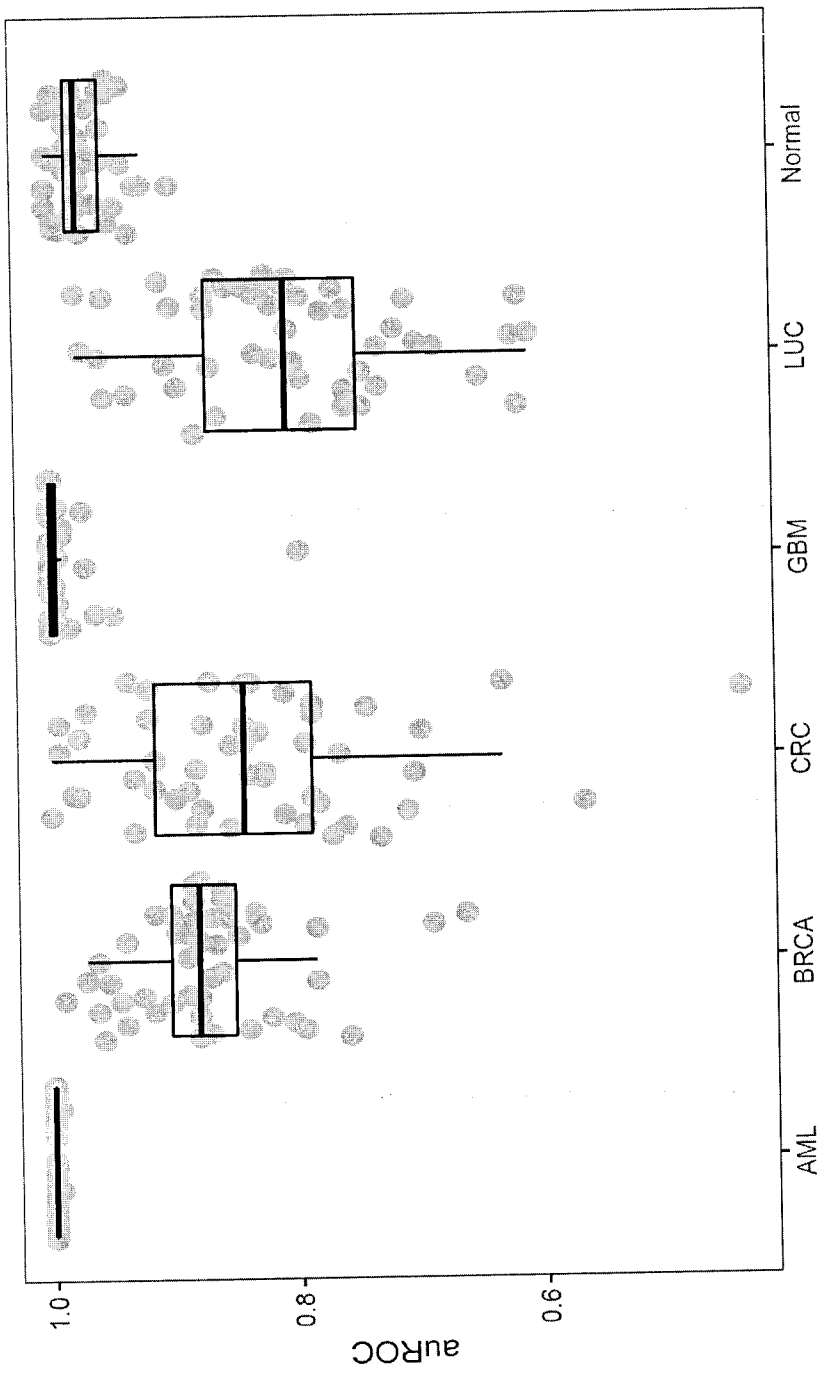


FIGURE 2D

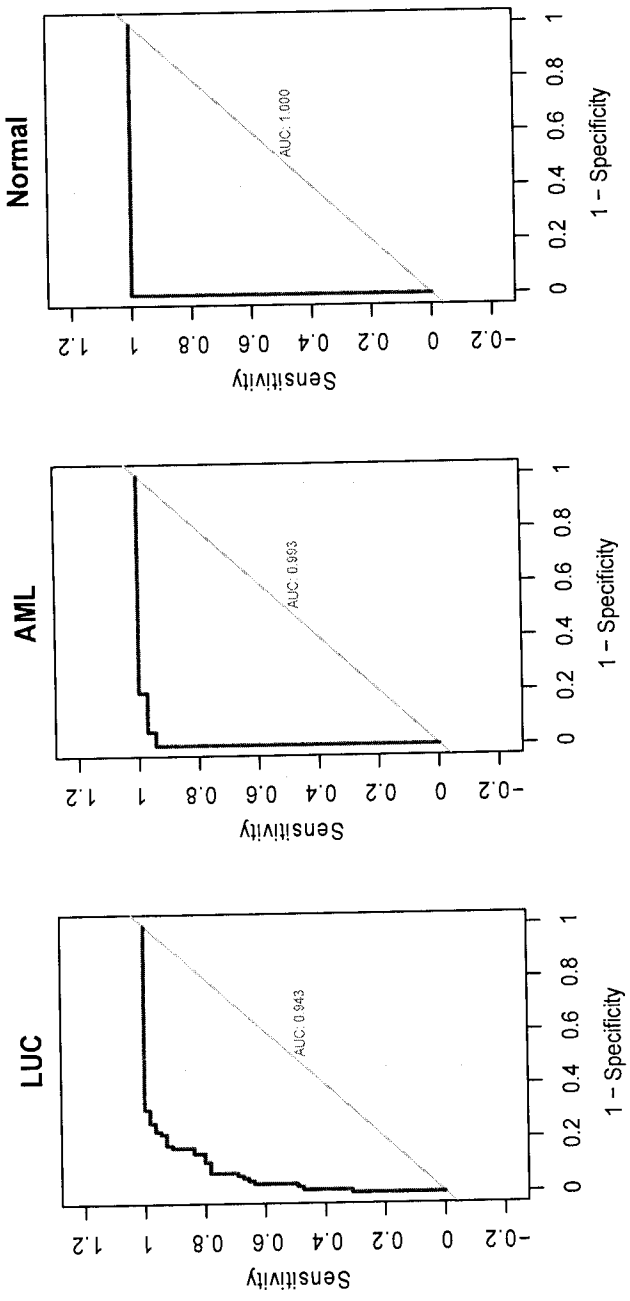

Normal
AUC: 1.000
Sensitivity
1 – Specificity
AML
AUC: 0.993
Sensitivity
1 – Specificity
LUC
AUC: 0.943
Sensitivity
1 – Specificity

FIGURE 3A

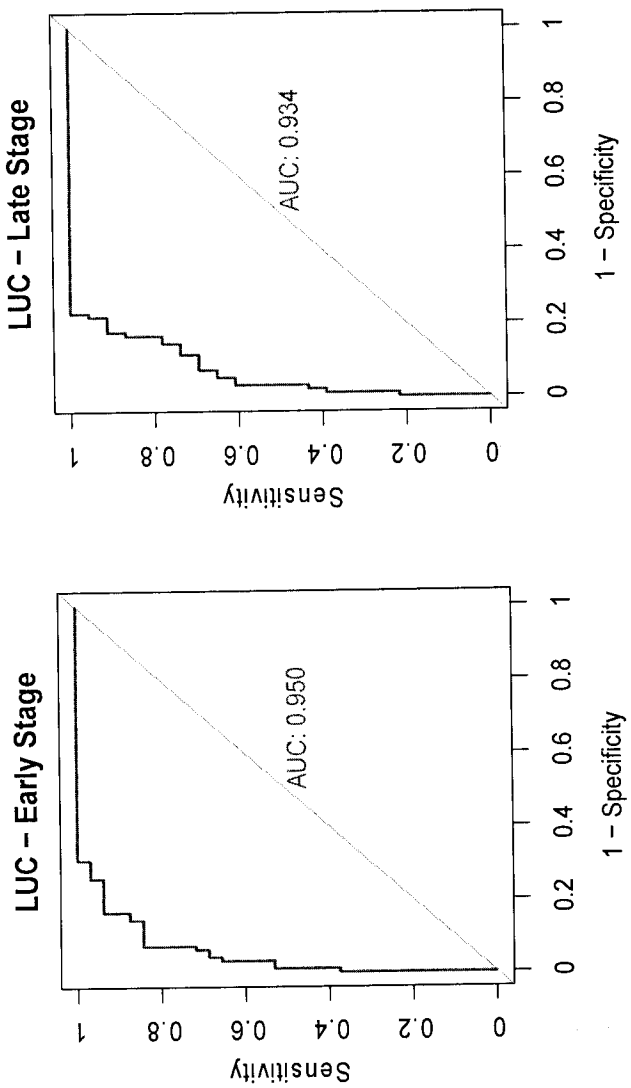

LUC – Late Stage
AUC: 0.934
Sensitivity
1 – Specificity
0
0.2
0.4
0.6
0.8
1
LUC – Early Stage
AUC: 0.950
Sensitivity
1 – Specificity
0
0.2
0.4
0.6
0.8
1

FIGURE 3B

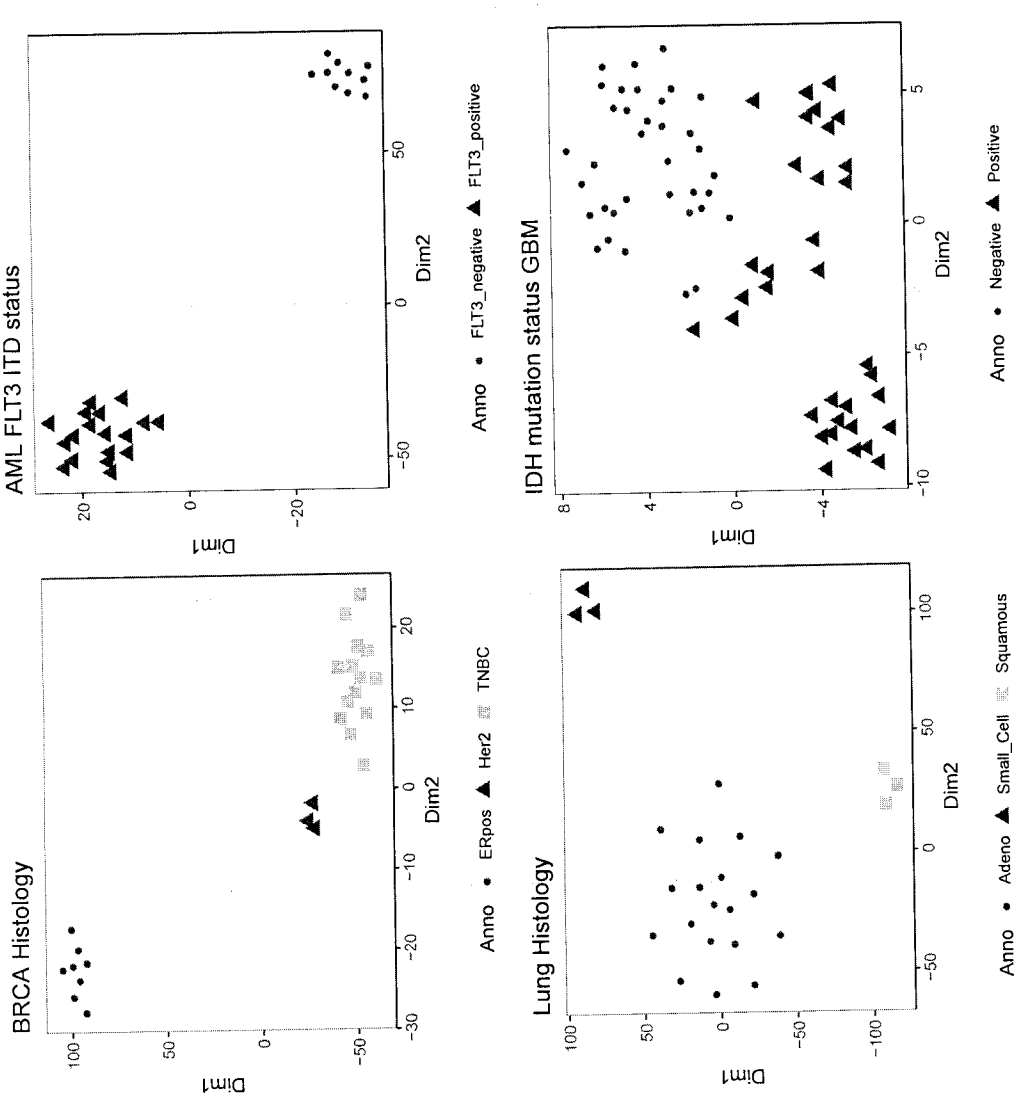

BRCA Histology
Dim1
Dim2
Anno ERpos Her2 TNBC
AML FLT3 ITD status
Anno FLT3_negative FLT3_positive
IDH mutation status GBM
Anno Negative Positive
Lung Histology
Anno Adeno Small_Cell Squamous

FIGURE 4A

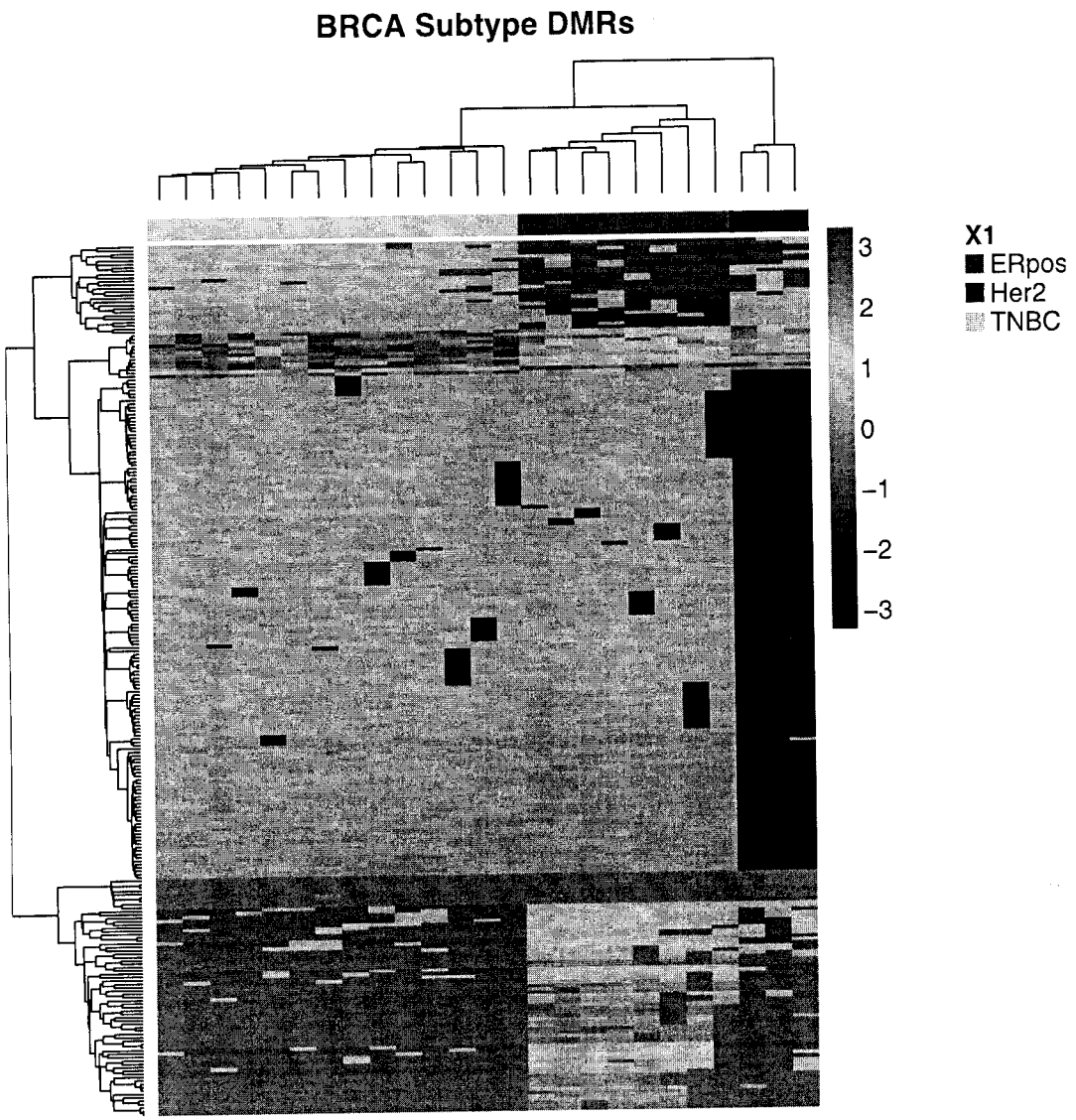


FIGURE 4B

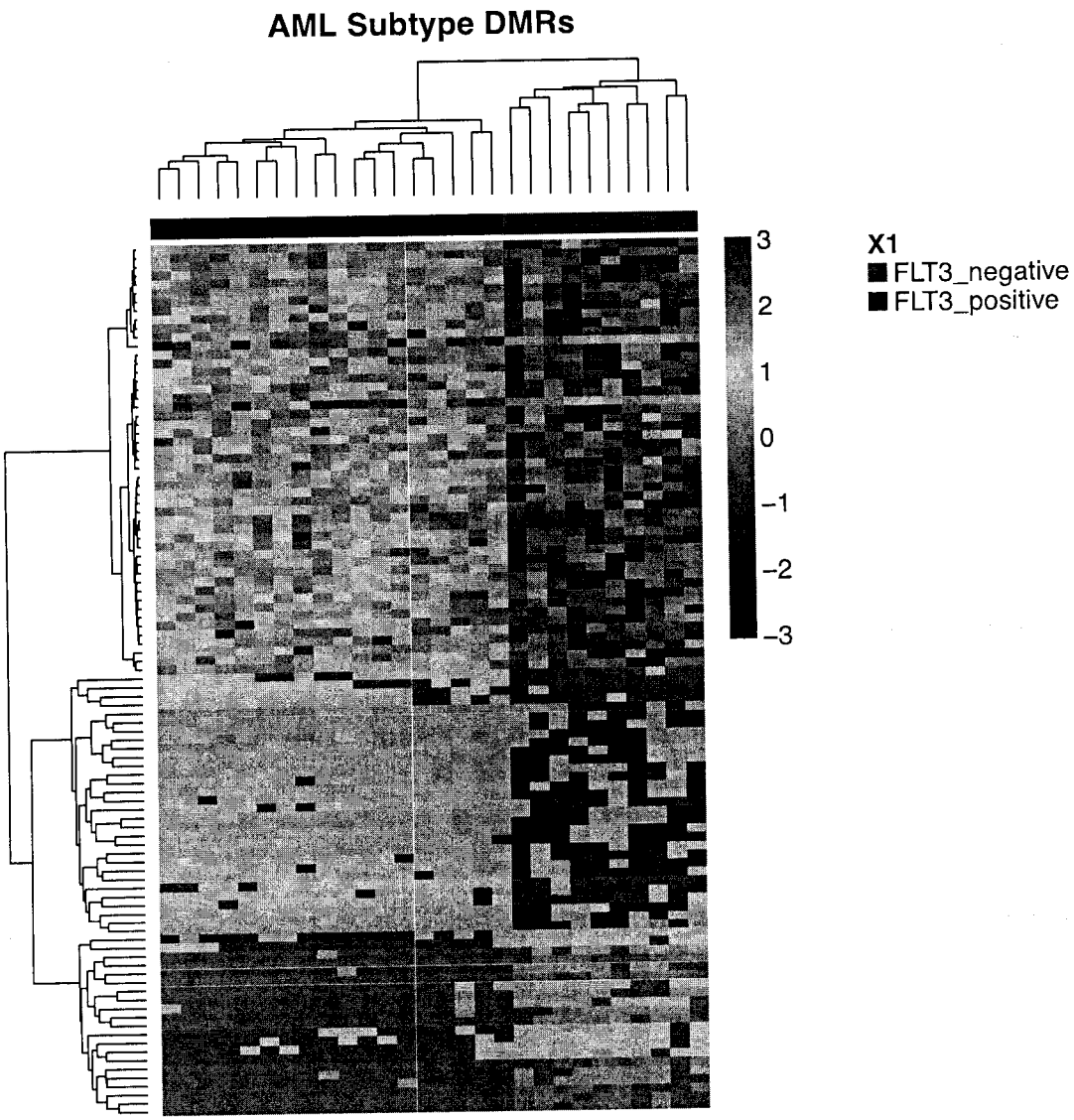


FIGURE 4C

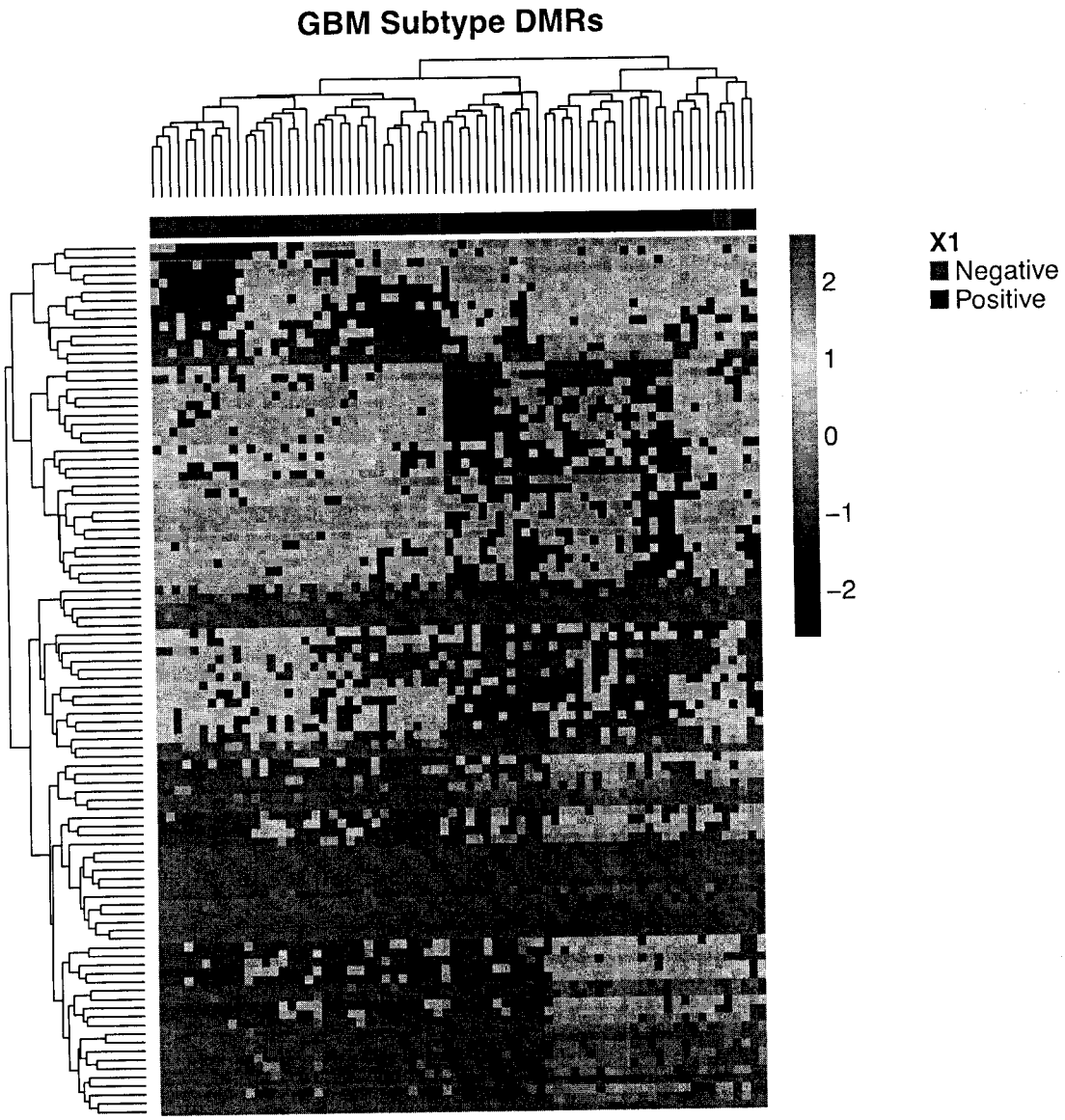


FIGURE 4D

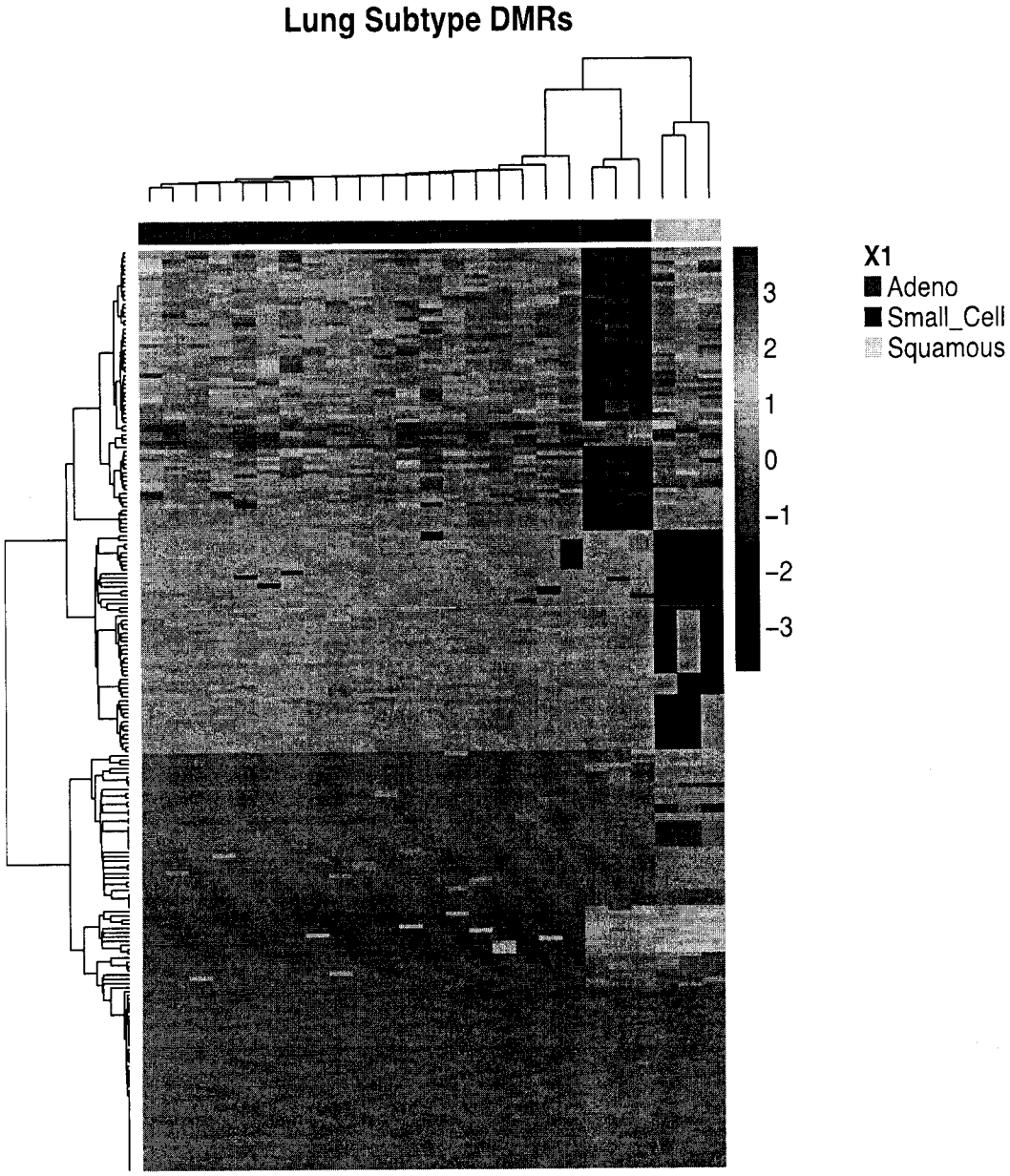


FIGURE 4E

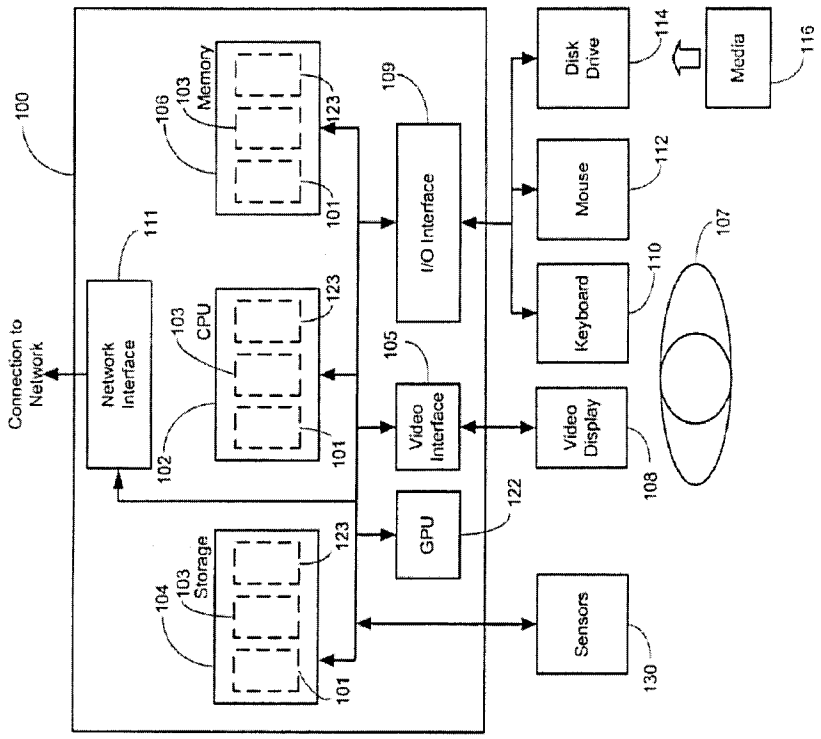

100
Connection to Network
Network Interface
111
102
CPU
103
101
123
Memory
106
Storage
104
GPU
122
I/O Interface
109
Video Interface
105
Disk Drive
114
Media
116
Mouse
112
Keyboard
110
Video Display
108
107
Sensors
130

Figure 5

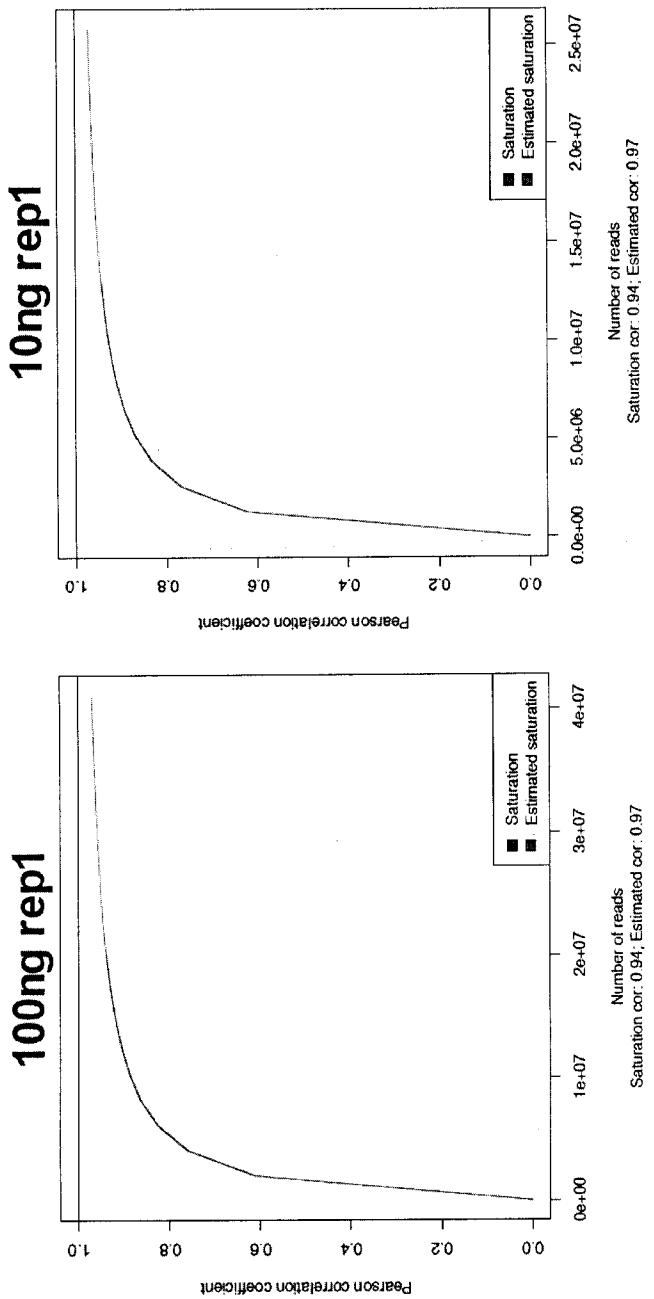

10ng rep1
Saturation
Estimated saturation
Pearson correlation coefficient
0.0
0.2
0.4
0.6
0.8
1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
Number of reads
Saturation cor: 0.94; Estimated cor: 0.97
100ng rep1
Saturation
Estimated saturation
Pearson correlation coefficient
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
1e+07
2e+07
3e+07
4e+07
Number of reads
Saturation cor: 0.94; Estimated cor: 0.97

Figure 6A

Figure 6A continued

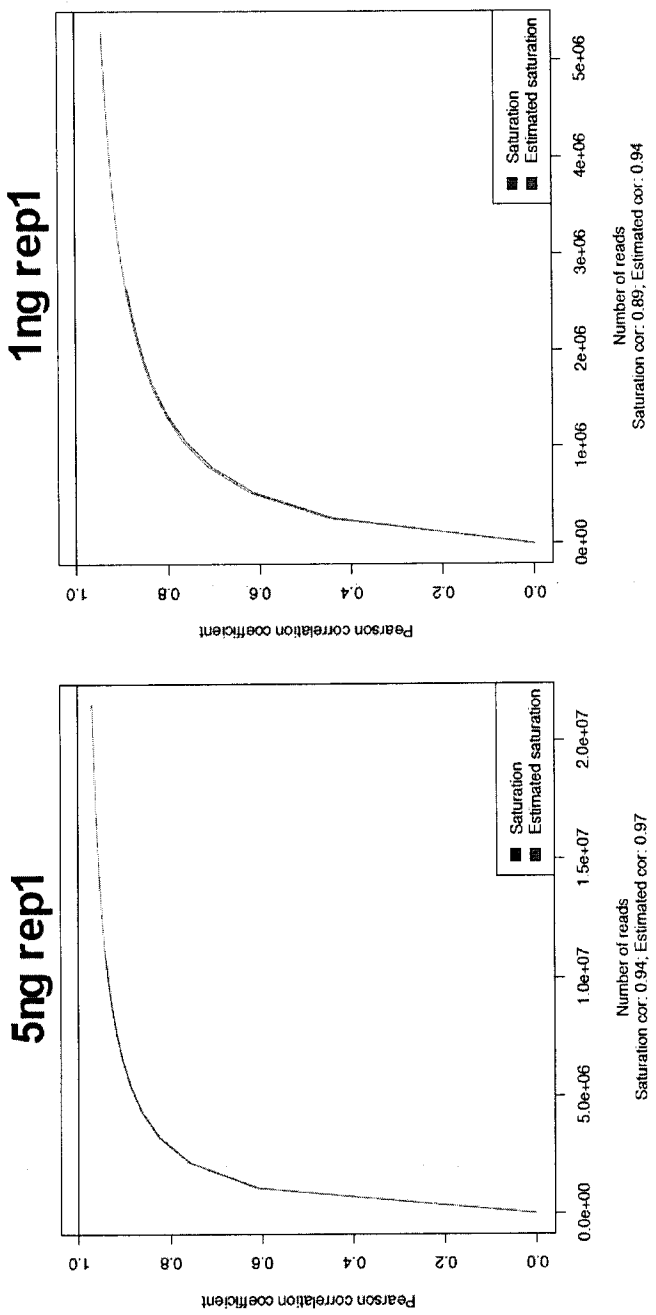


Figure 6A continued

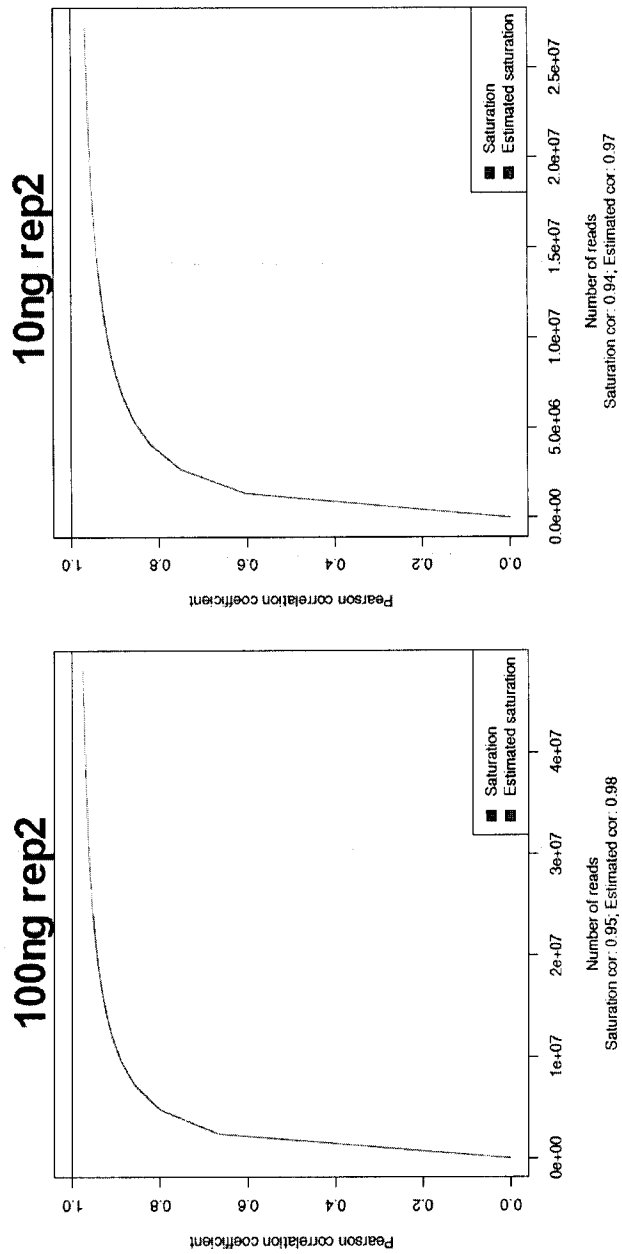
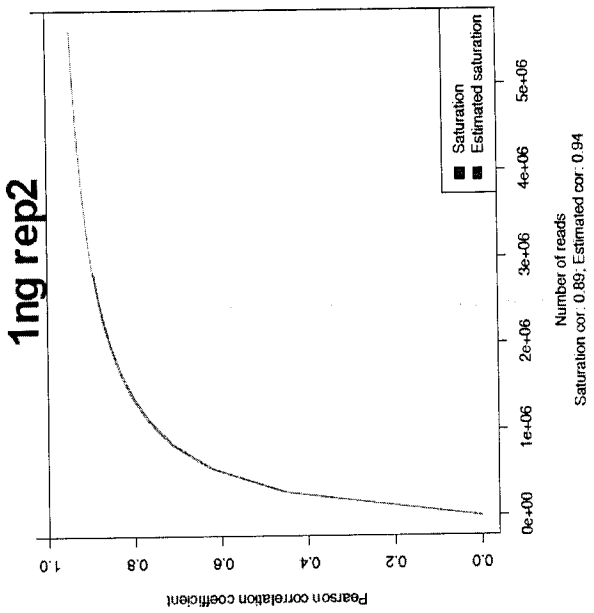


Figure 6A continued



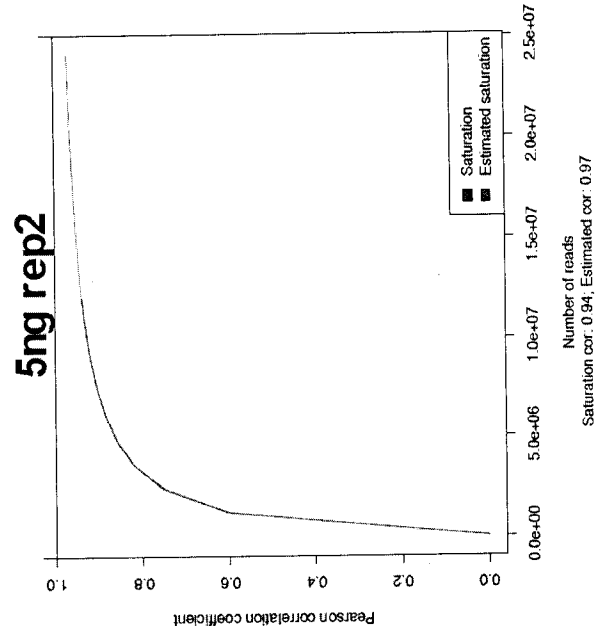
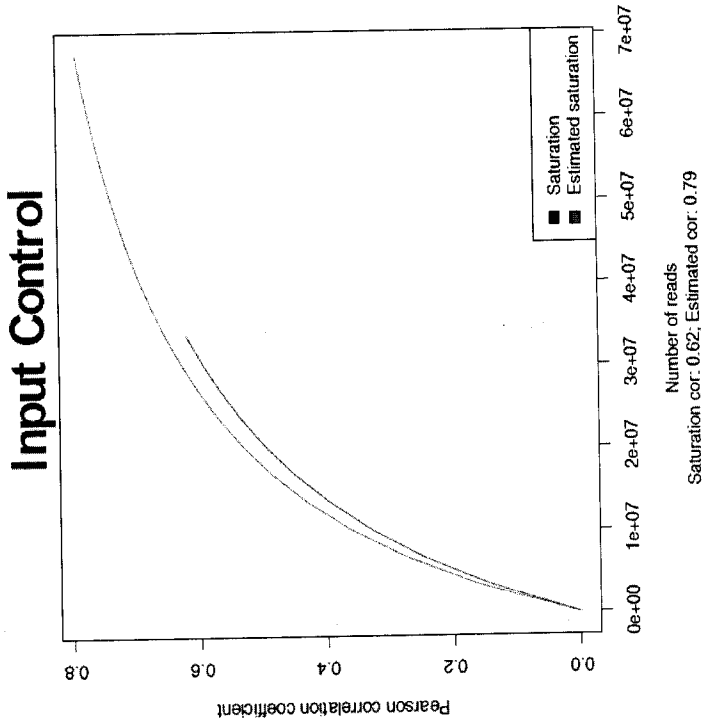
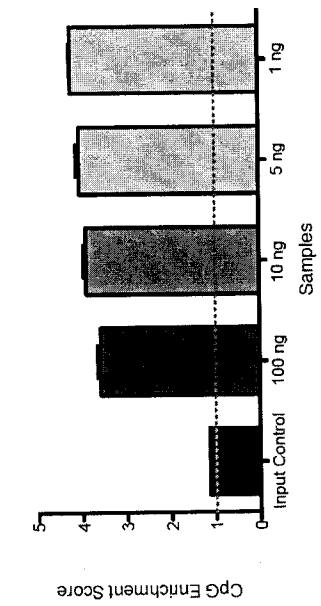
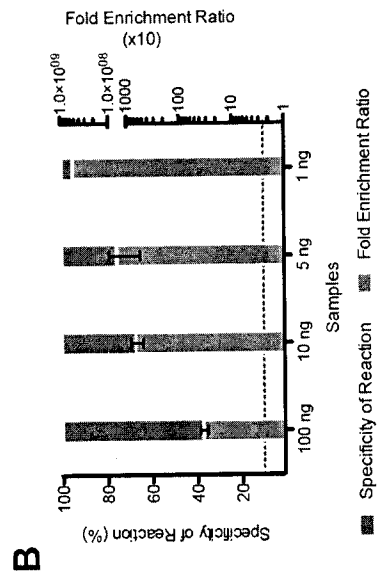


Figure 6A continued



Figures 6B and 6C





INTERNATIONAL SEARCH REPORT

International application No.
PCT/CA2018/000141

A. CLASSIFICATION OF SUBJECT MATTER
IPC: ***C12Q 1/6809*** (2018.01) , ***C12Q 1/68*** (2018.01) , ***C12Q 1/6886*** (2018.01) , ***G06F 19/22*** (2011.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Keywords searched across entire IPC

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic database(s) consulted during the international search (name of database(s) and, where practicable, search terms used)
Questel Orbit, Scopus, Google. Keywords: meDIP, meDIP-seq, methyl*, methylome, cancer, cell-free DNA, circulating tumor DNA

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	BUTCHER L.M. AND BECK S. (2015) "Nano-MeDIP-seq methylome analysis using low DNA concentrations" in: Haggarty P., Harrison K. (eds) Population Epigenetics. *METHODS IN MOLECULAR BIOLOGY* 1589: 115-138 Humana Press, New York, NY. ISBN: 978-1-4939-6901-2 *entire document*	1-38
Y	BURGENER J. ET AL. (2016) "Utilization of methylated circulating tumour DNA in oral squamous cell carcinoma for risk stratification and detection of recurrence" in: TFRI-Ontario Node Research Symposium Program & Abstracts. Accessed online: < URL: http://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=2ahUKEwi-k5nTyrrdAhUo6IMKHVjgD9wQFjACegQICBAC&url=http%3A%2F%2Fwww.tfri.ca%2Fdocs%2Fdefault-source%2Fdefault-document-library%2F2016-symposiumbook.pdf%3Fsfvrsn%3D6&usg=AOvVaw2jPwjGZr_UEpYfA38hvCQo > *entire document*	1-38

☑ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier application or patent but published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 13 September 2018 (13-09-2018)	Date of mailing of the international search report 03 October 2018 (03-10-2018)
Name and mailing address of the ISA/CA Canadian Intellectual Property Office Place du Portage I, C114 - 1st Floor, Box PCT 50 Victoria Street Gatineau, Quebec K1A 0C9 Facsimile No.: 819-953-2476	Authorized officer Mary Murphy (819) 639-6853

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	DE CARVALHO D. 16 May 2017 (16-05-2017) "Real-time liquid biopsy cancer diagnosis and monitoring" Flintbox. accessed online: <URL: https://www.flintbox.com/public/project/31470/ > *entire document*	1-38